Information Services & Use 41 (2021) 1–37
DOI 10.3233/ISU-210120
IOS Press

An Overview of the 2021 NISO Plus Conference: *Global connections and global conversations*

Bonnie Lawlor[*]
Guest Editor, NFAIS Honorary Fellow, 276 Upper Gulph Road, Radnor, PA, USA

Abstract. This paper offers an overview of the highlights of the 2021 NISO Plus Annual Conference that was held virtually from February 22–February 25, 2021. This was the second NISO Plus annual conference. The first one was held in 2020 and replaced what would have been the 62nd Annual NFAIS conference, but with the merger of NISO and NFAIS in June 2019 the conference was renamed NISO Plus and took on a new format. Little did they know that the second conference would have to be held virtually while the world was battling a global pandemic. The 2021 audience represented a 400% increase over the 2020 in-person attendance. There was no general theme, but there was a topic for everyone working in the information ecosystem - from the practical subjects of standards and metadata quality to preprints to information privacy and ultimately to the impact of Artificial Intelligence/Machine Learning on scholarly communication. With speakers from around the world and across time zones and continents, it was truly a global conversation!

Keywords: Linked data, standards, digital humanities, NISO, CRediT taxonomy, tenzing, metadata, artificial intelligence, machine learning, discovery services, FAIR data, privacy law, metrics, open access, indigenous knowledge, digital preservation, preprint servers, analytics, libraries, misinformation, fake news, knowledgebases, information sharing, COUNTER, information discovery, Miles conrad lecture, scholarly communication, historical bias, monopoly, surveillance capitalism

1. Introduction

In February 2020 NISO held the first NISO Plus Annual Conference in Baltimore, MD, USA. It replaced what would have been the 62nd Annual NFAIS conference, but with the merger of NISO and NFAIS in June 2019 the conference was renamed NISO Plus and took on a new format. The goal was to continue some of the best traditions of past NFAIS conferences while incorporating plenty of time for discussions and hopefully identifying areas where NISO could work to develop new standards and recommendations to create greater efficiencies in the information community. The inaugural conference was labeled a "Grand Experiment" by Todd Carpenter, NISO Executive Director, in his opening remarks. When he closed the conference, all agreed that the experiment had been a success (myself included), but that lessons had been learned and in 2021 the experiment would continue. Little did we know that for most of us in attendance it would be the last in-person conference in which we would be able to participate for more than a year due to COVID-19.

Fast forward one year and the second NISO Plus Annual Conference was held in a completely virtual format. The organizers took advantage of a very difficult situation and used it to create a unique and innovative conference. Faced with the requirement for virtual participation by both speakers and attendees

[*] E-mail: chescot@aol.com.

they chose the theme "global connections/global discussions" and recruited speakers from around the globe - many of whom would not have had the time or ability to come to the USA to speak. The planning committee itself was composed of twenty-seven people from eleven different countries. The conference attracted more than eight hundred attendees - an increase of 400% over 2020 with 21% of them being outside of the USA! Participants came from twenty-six different countries, nearly every inhabited time zone, and from every continent except Antarctica. They were representative of the information community, for example, librarians, publishers, system vendors, product managers, technical staff, etc., and from all market segments - government, academia, industry, whether for-profit or non-profit.

Todd Carpenter noted in his welcoming remarks that NISO wanted to make the conference experience about the attendees, their organizations, and their problems. He reiterated that the tagline for the conference is global connections and global conversations and that those terms summarize what they hoped to accomplish with the program. While it is not the same as being together in person, he noted that they structured the sessions with the objective of fostering real discussions. The topic of each session was to be introduced by a short set of talks followed by in-depth discussion at the end of which the moderators would note any concrete project ideas that came up during the discussion as potential projects - anything from improving metadata for the preservation of indigenous knowledge to improving testing methods for accessibility, or perhaps developing identifiers for packages of content. Attendees will be asked to volunteer for projects that are a match for their interests and expertise.

Carpenter was open and realistic in saying that not every session will generate a project, nor will every project necessarily find a home in NISO. He added that NISO does not need to be the home for every project idea generated at NISO Plus and that those taking the lead with the ideas are free to take them to other organizations such as the American Library Association, the Research Data Alliance, the new STM Innovations Lab, etc. He said to think of the conference as an incubator of tech ideas for the information industry. NISO's goal is to see that hundreds of seeds are scattered - some of which will sprout and others that will not. But hopefully one or two will grow over time into something that changes the information landscape in a significant way. We will only know if NISO Plus is a real success in perhaps three, five or ten years. He then thanked the NISO staff for their work behind the scenes and then everyone was given an overview of how the event would proceed from a technical perspective (it went *extremely* well).

I can attest that at least for the sessions to which I listened the discussions were interesting, in-depth, and some did generate ideas. Be forewarned - as I noted last year, I am pretty good at multi-tasking, but even I cannot attend two meetings simultaneously and I did not after the fact listen to every recorded discussion. In fact, not every talk was recorded – it was done at the speakers' discretion. Therefore, this overview does not cover all the sessions, but it will provide a glimpse of what transpired and perhaps motivate you to attend next year's meeting. I want to note that NISO has made the content of the conference openly available, so if something interests you, you can access a video of the presentation [1].

2. Opening keynote

The Opening Keynote was given by Cory Doctorow, a science fiction author, activist, and journalist [2]. The focus of his presentation was the digital manipulation of society, primarily by Big Tech companies as noted by Dr. Shoshana Zuboff in her book, *The Age of Surveillance Capitalism*: *The Fight for a Human Future at the New Frontier of Power* [3]. In Zuboff's research, she claims that the two companies, Google and Facebook, gather very large numbers of data points about their users, with the core purpose of making profit. Surveillance Capitalism also can be used to improve the targeting of political advertising to maximize its impact on the electorate.

Doctorow also referenced a Netflix documentary, *The Social Dilemma* [4], that looks at how the Internet's most popular products work on a basic business model of tracking users' behavior to sell targeted ads and induce addiction in a vicious cycle. The film discusses several issues, including how tech companies have influenced elections, ethnic violence, and rates of depression and suicide.

He said that there is much that he disagrees with in the Surveillance Capitalism hypothesis, but he does agree that Big Tech controls our lives and that is not good. Our world and our lives are better when we can exercise self-determination - find out several options for our lives and choose the one that suits us best. Where he parts ways with the hypothesis is on this point: "Surveillance Capitalism says that companies control our lives through highly-automated data-mining tools that allow them to manipulate us for money".

Doctorow believes that companies control our lives through monopoly, not mind-control. He said that if the latter works at all, the effects are short-lived. He referred to a Facebook experiment to impact voting [5]. About three hundred and forty thousand extra people turned out to vote in the 2010 U.S. congressional elections because of a single election-day Facebook message, estimate researchers who ran an experiment involving sixty-one million users of the social network. He believes that Facebook released information on the experiment because it thought that it proved that it had a mind control ray to which it could sell access - a self-serving claim thinly-supported by data that the company itself gathered. But these claims to mind-control are self-serving, and critics who repeat them are helping the tech industry sell its core product - high-priced, unfalsifiable attempts to manipulate users. He noted that Big Tech's defenders and detractors are apt to counter that the fact that their advertising makes billions in revenue means that it must work. He stated that this is no proof at all, adding that hedge fund managers consistently underperform simple index funds, but that they still rake in trillions of dollars from the most sophisticated investors who could make more money if they just put their cash in a no-load Vanguard fund. Similarly, techies and their bosses - some of the highest paid people on earth - are convinced that they are good at bypassing users' cognition to influence their behavior. These techies are convinced that they can change user behavior; convinced that it is moral for them to do so, and deluded about what is really going on. What is really going on is the same thing that's been going on since the gilded age - Monopolies!

2.1. Control through monopoly

Doctorow then went on to talk about digital interoperability and that monopolies want to prohibit complete compatibility between competitors (think Apple vs. Microsoft and "Office for the MAC - not as seamless as a user would like). He said that monopolies are a form of corruption.

Therefore, it matters whether the problem with Big Tech is because they are monopolies or because they mind-control. If the tech companies have created an immortal superweapon that strips us of our free will, then it would be catastrophic to break up those companies and distribute their mind-cannons into more hands than we could ever hope to enumerate, regulate, or tame. But if it is because they are monopolies, then Doctorow believes that we must shatter their concentrated power as a means of restoring good governance.

Rather than looking to exotic explanations for market concentration such as network effects [6], Doctorow said that we need to look at how we enforce monopoly law. After all, it's not network effects, or first mover advantages, or data, that resulted in the USA having the following monopolies: one eyewear company; two beer companies; three record labels; four movie studios; and five publishers (soon to be four if Random House's purchase of Simon & Schuster gets regulatory approval - it did, after the conference in May 2021) [7]. How did this happen? He said that Robert Bork [8] changed the face of competition law because he painted monopoly as a purely consumer issue and said that monopolies should only be fought when there was "consumer harm". Bork purported that unless it can be proven that a monopolistic course of conduct would immediately lead to higher prices, the monopoly should be allowed to form. Doctorow said to be realistic, it is nearly impossible to prove beyond a doubt that prices will rise after a merger. As a result, monopoly formations are rarely blocked, even though the overwhelming evidence is that mergers, acquisitions, and vertical monopolies eventually raise prices. He added that the fundamental problem of "consumer harm" is in the phrase - it conceives of us as consumers, not at citizens. Monopolies are bad not because of "consumer harm", but because of democratic ones.

While Doctorow made the argument that the problem is due to monopoly, not a mind ray, he made it clear that he believes that the Big Tech monopolies do take away our free will. For example, they prohibit us from installing apps of our choosing unless those apps are approved by their central committees; they tell us which ink we can use in our printers, which parts we can put in our phones, and which repair depots can fix our devices. These are huge impositions on free will and self-determination. He said that there is a saying that if you are not paying for the product, you are the product [9]. He continued, noting that the farmers who aren't allowed to fix their own John Deere tractors did not get those tractors for free. And no one is given a free iPhone in exchange for promising only to get it repaired at a specific place. The thing that determines whether a company sees us as a product is whether they can get away with treating us like a product. The lack of competition and anticompetitive laws such as software patents, anticircumvention, and enforceable terms of service deprive us of the choice to punish them for mistreating us.

Doctorow said that he wanted to close his presentation by talking about ideology - specifically, what are the ideological factors that make the Surveillance Capitalism hypothesis so robust? He thinks that the success of Surveillance Capitalism is the result of a strange ideological marriage between three groups of people who want to preserve the status quo for their own reasons. The first is the Big Tech industry, which is invested in its own status as a bunch of geniuses. If they are labeled as evil geniuses, then at least they get to hold onto their genius status. The second group is the true believers of capitalism, who practice a form of tech exceptionalism when they paint tech as a "rogue capitalism" whose mind-control rays short circuit the market's near-mystical ability to self-correct. He said that if tech is a rogue capitalism, then the problem is tech, not capitalism. And finally, the third group is comprised of the people who are thriving under the status quo. And to those groups of people, he said the following: (1) Big Tech did not conquer the world through genius, but through the same mediocre sociopathy practiced by the Rockefellers, Carnegies, Bells, and others; and (2) Markets tend towards monopoly, and our minimum response should be strong state intervention that prohibits anticompetitive mergers, restricts corporate power, and punishes anticompetitive conduct. Surveillance Capitalism is not a rogue capitalism - it is simply capitalism. It is three sociopaths in a trench coat. We have dealt with them before, and it is long past time that we dealt with them again!

3. Linked data and the future of information sharing

3.1.

The first session that I attended after the opening keynote focused on the importance of linked data and the challenges that still need to be overcome. The initial speaker was Christian Herzog, Co-Founder and CEO of Dimensions [10] and the Chief Portfolio Officer of Digital Science [11]. He opened by saying that he was not using the term "linked data" in its purest technical form; he uses the term to describe the connections and relations that are of use to the end user. Basically "linked" is seen from the angle of the consumer of the data, not from the engineer who builds and provides the data infrastructure. He noted that a lot of players are involved in producing the data, so a lot of agreements need to be reached as to how the data should be linked, how it should be provided, and which identifiers should be used. But afterwards, the data also needs to be in a certain infrastructure so that the user can have the confidence to jump into the analysis.

Herzog noted that Dimensions is part of Digital Science whose vision is of a future where a trusted and collaborative research ecosystem drives progress for all. Within Digital Science, Dimension plays the role of enabling the linking of data from "idea to impact". They have created the world's largest linked information dataset - covering millions of research publications that are connected by 1.5 billion citations, supporting grants, datasets, clinical trials, patents, and policy documents. It is the most comprehensive research grants database that links grants to the resultant outputs of the funded research - millions of publications, clinical trials, and patents. It includes datasets from repositories such as Figshare, Dryad, Zenodo, etc. The dataset also covers policy documents with more than 2.1million links to publications that demonstrate the societal impact of scholarly work.

The company invests significant effort to make diverse data silos as interlinked as possible. They have also created a user interface that is easy to use - one need not be a tech expert. Herzog also noted that they released the Dimensions data on Google BigQuery because there they found a solution as to how they could make the underlying data available to every user in a large relational database. Google Big Query provides the computational infrastructure and power that allows a user to jump into data analysis right off the bat by creating an account and starting, even without learning new skills if they happen to be able to use SQL queries already or if they connect it to standard Business Intelligence tools. I found an interesting on-demand free webinar that walks you through Dimensions on Google BigQuery [12]. The webinar is also available on the Dimensions website. I should note that the dataset is freely-available for personal, non-commercial use. I have not taken the time to play with it, but I will. It looks impressive.

3.2. Cite data. Link data

The second speaker was Shelley Stall, Senior Director, Data Leadership, at the American Geophysical Union (AGU). The title of her presentation was "Cite Data. Link Data". She opened by saying that there is quite a lot of things that can be done to make sure that data is well-documented, preserved, and reusable, but one of the most critical things that a researcher needs to do is to make sure that the data is in their paper is cited, and then linked into all of the other research objects coming through from their research and the work of their colleagues. She put up a slide that displayed an interactive diagram that was done by the journal, *Nature*, to celebrate one hundred and fifty years of publishing from 1869 through 2019 [13]. The diagram showed how the papers are connected based on the references from one paper to the other. And the interactive nature shows you, like based on seminal papers, what it was based on and then what papers came from it over time. The diagram also shows the interconnectedness of diverse scientific disciplines and how the discovery that takes place in one discipline can impact the work of others.

She then asked a "What if" question. What if we took that underlying data for the diagram and made it even more connected? What if we were able to have that structure immediately? We could see the connected data sets, and the connected software, the models, know if there was a clinical trial, etc. And we could take this further. Who are the authors of the paper? What are their affiliations? Wen would be able to see every single research product that came from all the researchers affiliated with a particular college or international effort. That is exciting (when you look at Dimensions you can see that this is already being done).

What do you need to create these links? Well, you need at least two entities. For a publisher such as AGU, the publication itself is usually one of the primary ones, but that is not really required. It could be the data that comes from software. It could be the developer for that software, etc. And then to connect these entities uniquely and accurately a persistent identifier (PID is required. Then what is the relationship between the entities? Does each one cite the other? Is one cited by the other? And then this needs to be machine readable in a way that is consistent across all researchers. It is not hard work, but it does require upfront thinking.

Stall then showed some slides that were created by Helena Cousijn from DataCite who had recently given a talk on sharing and citing data. DataCite [14] is a registration entity that creates digital object identifiers (DOIs) that serve as persistent identifiers for data and other things. When a DOI is registered, they track how it's connected. The workflow actually is complicated and involves the following: (1) Journal policies requiring a citation; (2) the author selecting an appropriate repository for that data; (3) making sure that there is coordination between getting the paper and getting the data preserved; (4) linking it all; (5) getting it published correctly with all of the right persistent identifiers in the right place, and (6) getting them distributed, and aggregated, and then making everything available for others to find it.

Most publishers require ORCID [15] iDs for an article's primary author and Stall highly recommends that publishers require them for coauthors as well. We know what the institutional affiliations are. And many publishers are starting to require data citations for datasets and, where it is appropriate, citations for software (see an article on software citations by Daniel Katz and Shelley Stall that is elsewhere in this issue of *Information Services and Use).* There also are organizational identifiers - ROR [16] and GRID [17] who work closely together. There is also FREYA that was started by the European Commission in December 2017 and ended in December 2020 [18]. It aimed to build the infrastructure for persistent identifiers as a core component of Open Science, both in the European Union and globally. The latter initiative created the concept of a PID graph which is a means to take those links and relationships and display them in a visually- useful way. You can make inferences not only between two entities, but you can also infer things across links between three entities. Using the PID graph you can take a reference in a paper and explore what other connections that it may have. Prior to this tool being developed, it was difficult to see a data set. She said that AGU knew that they had referenced them, but it was hard across all repositories to see what was being registered in a particular repository in a way that it was connected to other research outputs. Stall added that it is important to realize that the authors of a specific paper are not always the creators of the data sets that were used in the research, and credit needs to be given to those who have generated that data.

She went on to say that everything may sound great, but the system does not work that well. AGU has partnered with the Global Biodiversity Information Facility (GBIF) [19] for a long time. GBIF has done an automated and manual review of every single paper in their disciplines to determine if authors were citing data and doing so according to GBIF guidelines. What they found was that authors have been increasingly citing datasets since 2018, but that the number of citations that do not follow the guidelines still significantly outnumber the number that are compliant (69% of the 2020 citations were not in the correct format!). Even with a lot of work in communicating with authors on what the citations should be, publishers still have a lot of trouble helping authors make those citations correctly. And this is one of the areas that she would like everyone to work on together. Certainly, AGU faces several challenges and these numbers from GBIF demonstrate that we all still have a lot of work to do. We need: (1) journal policies that require data citations and the use of persistent identifiers; (2) citation validation/copy-editing that is specific to data (or software) because data and software citations are formatted differently than journal and book citations and we need to make sure that they are machine readable; (3) to be aware that there are PIDs being used that are not DOIs; (4) to be aware that PIDs are being registered with a variety of registration agencies and that not all are registered with Crossref; (5) to be aware that there are URLs being used that are not PIDs, but do have valid locations for where that data and software are located because of how repositories are evolving. Not all repositories have the PIDs that are needed, but those repositories might actually be the right repository for that data; (6) active involvement with Crossref and the new schema that will be released this year (2021); and (7) to provide authors with examples of well-cited data sets and software packages. She urged everyone to take a close look at how they are validating citations, and to make sure that the machine-readable XML entries that go to Crossref to populate the links for the PIDs are actually accurate.

In closing, Stall recommended that attendees listen-in on another session entitled “Research data: describing, sharing, protecting, saving” that was to be held that evening to hear about what is in the works for the new Crossref schema that will allow publishers to do a better job identifying data and software citations. For convenience, a summary of the Crossref portion of that session follows. Also, as I mentioned earlier, an article on software citations by Daniel Katz and Shelley Stall appears elsewhere in this issue of *Information Services and Use*.

4. Research data: Describing, sharing, protecting, saving

4.1.

Patricia Sweeney, Head of Metadata at Crossref [20], spoke about Crossref’s plans to help publishers to get better citations for data sets. She noted that citations are a core part of our infrastructure, so for most publishers, Crossref metadata is really the key to having their citations identified and connected to research. Crossref has been on a long road to handling citations efficiently and we want to make this easy. Conceptually, it is very easy. Crossref members gather citations of all kinds, including those for data and software, and then those citations are sent to us. And then we pass them along the chain. But if it is that simple, why isn’t happening? She said that unfortunately, there are a lot of “ifs” involved in the process. If members send them a citation, they do pass it along to the Crossref REST API outputs and XML outputs. But identifying whether a citation is a data citation or a software citation and what should be done with it can be *very* tricky. The first step depends on members collecting data citations in the first place and understanding how and why those should be supplied to Crossref. That involves technical and cultural changes, and both are very hard. Crossref has been discussing data citations with their membership a lot over the past few years. But, unfortunately, our support and guidance for this has not been robust, partially because citation practices community-wide have not been clearly defined for a long time.

However, she believes that while this has changed recently, practices have been evolving. And more importantly, she noted that Crossref has their own limitations as to what changes could be made. The organization is emerging from a heavy load of technical debt and has not been as nimble as they would like. But hopefully this will change. For a while the organization was making recommendations that worked within what they could actually support, not really in the way that they would ideally like to support data and software citations. She admitted that the recommendations were not clear or easy for their members to follow, and as a result there wasn’t much uptake. But for those who are sending Crossref data citations as Shelley mentioned, they match the citations to the DOIs and that works great for journal articles. But it does not work well if an item does not have a DOI, as is the case for many software citations and even data citations, or if the citation for does not have a Crossref-specific DOI.

Members can supply Crossref with a citation that has a DataCite DOI. And if they supply the DataCite DOI and explicitly say that it is a DOI, Crossref will be able to pass that along. But if a member does not have the DataCite DOI included in their citation, for technical reasons Crossref cannot add it in for them. Unfortunately, many publishers do not collect those DOIs nor do any matching on their end. As a result, the data citations get lost. She noted that for step three, citations are available by Crossref APIs, and those citations are passed along. Crossref does send all citations that members send to them and opt to make public to their JSON and XML outputs. And they are available for downstream users. And those also include any user interface matches that Crossref can make. But data and software events are in their Event Data API and the Scholix API as well. And those two are key to making the connections between data and software downstream. This means that if Crossref does not have a DataCite DOI for a data citation, they cannot send it to their Event Data API or to the Scholix API that are essential to connect data and software to research. This is a weak link in the flow.

She then provided a few examples. Her first example was a data citation that includes a DOI:

```
<citation key = “ref3”>
<unstructured citation>Morinha F, Dávila JA, Estela B, Cabral JA, Frías O, González JL, Travassos P,
  Carvalho D, Milá B, Blanco G (2017) Data from: Extreme genetic structure in a social bird species
  despite high dispersal capacity. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.684v0
</unstructured_citation> <doi>10.5061/dryad.684v0</doi>
</citation>
```

This can be sent where it needs to go – XML and REST API, Event Data, and Scholix. However, the same citation without a DOI as follows:

```
<citation key = “ref3”>
<unstructured citation>Morinha Morinha F, Dávila JA, Estela B, Cabral JA, Frías O, González JL,
  Travassos P, Carvalho D, Milá B, Blanco G (2017) Data from: Extreme genetic structure in a social
  bird species despite high dispersal capacity. Dryad Digital Repository. </unstructured_citation>
</citation>
```

will get passed along to Crossref’s XML and JSON outputs, but *not* to Event Data nor to Scholix. Anyone looking specifically for DOI citations will miss this reference. It is lost in a blackhole.

She also displayed a well-formatted software citation. She noted that it has a software-specific identifier, but added that Crossref does not recognize it as being a software citation. They recognize it as being as a citation that does not have a DOI attached to it, but has an identifier with which they do nothing. This citation also gets lost in the blackhole. She said that this illustrates what Crossref is dealing with now.

However, they are making some changes to allow data and software to be identified and passed along. And these changes are simple. They will allow members to flag citation type - data or software - and to find identifiers in their citations. And for those who supply structured references, they can add some data-specific pieces of metadata. She is hoping that these changes can easily be added to their members’ workflows, particularly for those who are collecting data and software citations already. And for the citation types, Crossref is adding data as a citation type, software as a citation type. And they are also adding journal article, book, preprint, etc. as a citation type. She added that if Crossref can get their members on board, these changes will really make the reference metadata a lot more useful. She added that since they are aligning with the JATS recommendation [21] that this is something that members can do without too much pain. This will allow citations to be more specific (data, software, book, article, preprint, etc.) in a way that Crossref can work with as can anyone downstream who uses the Crossref data.

She closed by saying that these changes are not in place, but that they hope to have this up and running in a few months. They are not far enough along to be able to predict timelines, but hopefully in the spring of 2021 there should be some news (*Note:* I could not find any news to pass along).

4.2.

Rosemary Farmer from Wiley who is involved with the FORCE 11 Software Citation Working Group [22], closed this session with a brief statement that there are two key recommendations as to how to make sure that data and software citations are properly linked and that they get counted. These are first and foremost, for all stakeholders to ensure that their XML is consistent and aligned with industry standards for the most reliable output; and second, to agree to use persistent identifiers for software and data and establish validation mechanisms for them. Complying with these recommendations should help to ensure successful citation linking while being able to give credit to the creators. She closed by saying there are more policy and operational discussions and decisions that need to be made and that the FORCE11 Software Citation Working Group is continuing its work and will bring forth additional recommendations in the future. In the meantime, the group welcomes any input and questions.
Reminder: An article on software citations by Daniel Katz and Shelley Stall appears elsewhere in this issue of *Information Services and Use*. Also, you may be interested in reading some comments that Shelly Stall made at the NISO 2020 Conference regarding the problems with data citations as they flow through the process from publisher to through to Crossref and DataCite. The numbers that she presented serve to reinforce the issue [23].

5. Connecting the world through local indigenous knowledge

Tuesday morning opened with a keynote by Margaret Sraku-Lartey, Principal Librarian, CSIR-Forestry Research Institute of Ghana [24]. This was one of my favorite presentations because Sraku-Lartey opened my eyes to something I never really thought about. She said that scientists, information managers, and publishers have always been concerned and interested in explicit knowledge, perhaps because science is based upon empirical evidence that strives for objectivity, accuracy, and acceptability. All kinds of standards have been developed to ensure that scientific reportage is standardized, is of a high quality, and is able to stand the test of time. But she pointed out that little attention or interest has been paid to the tacit knowledge that abounds in our midst. She described this tacit knowledge as that which is garnered from personal experience and context, knowledge that is difficult to write down, to articulate, or to present in a tangible form. This is the vital knowledge that has accumulated over many generations - knowledge that is called Indigenous Knowledge (IK).

Sraku–Lartey focused upon the importance and value of local IK and how it is being threatened in today's modern world rather than being leveraged by the global information community to catalyze development. She specifically called out three types of IK: (1) medicinal knowledge related to human health, i.e., herbal medicine; (2) Sacred Groves - geographic areas set aside to preserve plants and animals and that can help to mitigate the impact of climate change; and (3) Living Libraries - communities of people who are also holders of cultural wisdom and history and who are custodians of all knowledge relating to the history of their own community. She gave examples of each and how/why they are important. She also noted how her organization is attempting to preserve IK.

They have started an initiative by digitizing local information, the local knowledge that people have come up with, and they now have a database of such local information collected from the field. When they come back to the Institute, they try to sift through the data and then examine the published literature to determine if the information that has been retrieved has been scientifically validated. If it has, they then digitize the validation article and add it to the database. For example, if there is a particular plant, they include the scientific name, the local name, the uses to which the people have put it and include the uses to which the published literature refers. This is one of the methodologies by which all stakeholders can integrate the knowledge from different sources, because once it is in a database, scientists are likely to look up the literature, see that it has been documented, and then use that as a starting point to go on with their own research. Researchers must authenticate the information and then do further research to make sure that it works for everyone. The Institute has started in a small way by developing a database for the plants, the food plants, and for local indigenous foods and medicines. She believes that if they can continue this effort that they could create a large database just as India has done for their own Indigenous Knowledge [25].

Sraku–Lartey added that the CSIR University of Science and Technology, which is one of the huge universities in Ghana, has an herbal medicine department in their pharmacy department where they integrate herbal medicine into orthodox medicine to see how it works. In fact, they train traditional pharmacists to use herbal medicine. This reinforcement cycle is a huge part of the scholarly research effort.

She closed with a call to action, requesting that scientists, librarians, publishers, and others in the information community collaborate and move forward together to save and build upon global Indigenous Knowledge.

Sraku-Lartey has written an article based upon her presentation that appears elsewhere in this issue of *Information Services and Use*. I highly recommend that you read it.

6. The future of intellectual property: AI and Machine Learning

6.1. Content can be used in AI/ML projects?

Roy Kaufman, Managing Director, Business Development, Copyright Clearance Center [26], opened this session with an excellent presentation on the use of copyrighted material to train computers for Machine Learning. He talked about the inputs to the process and noted that much of the input that goes into Artificial Intelligence (AI) I has nothing to do with copyright as it is raw data that is not protected by copyright law. His focus was limited to the use of copyrightable materials as input to train machines or used as input for AI. In this context, he lumped together books, journals, magazines, newspapers, etc. Kaufman noted that while copyright law may seem complicated, it is not. At its core, it is about the right to make copies for certain works that are protected. If a human reads a book, there's no copy being made. Copyright is not implicated in the process of human reading. However, for hundreds of years, machines could not read or use content. So, if you made a copy for a *human* to read, that implicated copyright law. And the reason he said this is we often talk about the right to read and the right to mine. But if the machine is "reading" a copy, you need to figure out what copyright law says about it. If a copy is not made, copyright doesn't apply. However, if a copy is made for a machine to read copyright law applies, and you must determine if making these copies is an infringement. Was it done under a license, which means that permission has been given by the rights holder? Or was it done under an exception or limitations, for example, some sort of excused copying for the greater good. He noted that he often runs into people who say that they know others who are interested in text mining and who believe that all input should be free. And he will ask are they with a software company or are they with one that s content? He made the following analogy. If you are making cars, you wish gasoline were free. You are not going to worry about the implications - you just want people to drive your car. Always determine when people are saying that content should be free where they are coming from. Everyone has a bias in this issue. He added that to take the analogy further if it is your software (the car analogy) your input better not be kerosene. The input must be fit for its purpose. It must be structured, refined, and appropriate for what is trying to be accomplished.

This is an area in the digital age that makes copyright law complicated. Kaufman said that all copyright really under international law should fall within the Berne Convention Article 9 Section 2, which states that there are exceptions for copying, i.e., there can be legal copying without the consent of the rights holder under special cases that do not conflict with normal exploitation. But what is normal exploitation? He noted that twenty years ago, using content for Machine Learning was not on the radar screen of most researchers. Normal exploitation today is when news companies will enter into a license with a hedge fund where all their data flows into a pipe and they do trades, and they make a lot of money on that. With text mining services humans do not always read the content. It is the machine that does the reading/sorting. The European Union (EU), which is more of a statutory regime, does not have a fair use analysis similar to that which exists in the United States. They have statutes that say you can or cannot do such and such. His opinion is that the EU has probably the most significant copyright law in the GDPR (*Not*e: this will be discussed in a later session).

What the EU has done is create a directive that is to be implemented in two years by every country, but he thinks that only Germany will meet the deadline. The EU created two copyright exceptions. One is for non-commercial research text and data mining (TDM). So, an educational institution that subscribes to a lot of journals can text mine if they are doing it for non-commercial purposes no matter what is stated in the license agreement. There is also a broader exception for everyone related to materials on the open web - it can be mined unless the rights holder opts out (which most publishers may do!). He noted that the EU is coming off a limited commercial TDM exception that exists in the United Kingdom. They are not identical, but to him they are similar, and it is a non-commercial exception. In the U.S., at least to date it has all been about fair use analysis.

He went on to say that people outside the U.S. think text mining in the U.S. is fair use. His response to them is, well, it's not a use. Making a copy to mine may very well be fair use. But the actual use isn't the mining. The use is what you're mining for. To provide examples he discussed two legal cases and something he termed a "non-case". The first example was the Google Books/Hathi Trust case. These are two separate cases, but they are similar, so he lumped them together. Regarding them, it was stated that making copies of print books to mine for non-commercial semantic research is likely "fair use". The text-mining example given in both the Google Books and Hathi Trust decisions [27] was non-profit linguistic research done specifically to determine when historically "the United States" began to take the singular ("is") instead of the plural ("are").

The "non-case" is that while Google started scanning the books in the library, they were also scanning journals in the library. They stopped scanning the journals because they found out that publishers were already scanning their works and selling them to libraries. He said that Google will never admit this, but they did stop scanning the journals for that reason. And it did seem to him as a lawyer who had a stake in this because he was working for one of the publishers involved, that it was also because their fair use claim was very weak. He said that he does not know for sure, but that is his opinion. The third case is more recent and while the phrase "text mining" was not used, it is similar.

This was a case involving TVEyes and the Fox News Network. According to the court documents, "TVEyes is a media-monitoring service that enables its subscribers to track when keywords or phrases of interest are uttered on the television or radio. To do this, TVEyes records the content of more than fourteen hundred television and radio stations, twenty-four hours a day, seven days a week. Using closed captions and speech-to-text technology, TVEyes records the entire content of television and radio broadcasts and creates a searchable database of that content. The database, with services running from it, is the cornerstone of the service TVEyes provides to its subscribers [28]". The initial decision was that what TVEyes was doing was "transformative" and therefore constitutes fair use. Fox appealed and the Second Circuit held that TVEyes was not protected by fair use because even though its work was transformative, its redistribution of Fox's copyrighted content "…deprived Fox of revenue that properly belongs to the copyright holder [29]".

This case ended up as an infringement, and Kaufman's point was that every case is fact determinate. That's how fair use works and therefore you cannot easily make blanket rules about fair use.

Kaufman closed by saying that if you need to license content for AI and text mining activities - remember that you can only license what is available to be licensed and this gets into issues of equity and accessibility. He posed four questions to consider regarding input for AI and ML activities

(1) Equity in Original Works: *Does the input adequately reflect voices of marginalized communities?*
(2) Equity in Licenses Availability: *What works are available for mining, and what biases went into the selection of those works?*
(3) Equity in Technological Availability: *What works are "use ready" in terms of tagging, formats, etc. What biases went into the selection of those works for investment*?
(4) Equity with Respect to the Content Creators: *In our efforts to be complete are we ignoring the voice of the author?*

He stressed that the output of AI is only as good as the input, and one seriously needs to avoid/minimize biases in the selection/creation of input. (*Note:* the issue of bias was discussed at length in a presentation by dana boyd at the 2020 NISO Plus conference and is definitely worth a read) [30].

6.2. Who owns the output of AI/ML content?

The second speaker in this session was Nancy Sims, Copyright Program Librarian, University of Minnesota, who talked at length about the reuse of material that people post on Facebook, Flickr, Tumblr. TicTok, Twitter, etc.

She noted that TikTok is a great example of a site where she believes that there is a fairly well understood cross-site expectation of reuse - of other people reusing your material and changing it. But she suspects there are even some TikTok creators who have some schisms in their perceptions of the right ethical way to use TikTok. Another example is on Twitter. From her experience she has learned from seeing other people's use of thread reader apps that if someone posts a long thread, a bot can be invoked at the end of the thread that will take all the tweets in the thread and post them - usually as a separate web page - where the tweets can be read all together. A lot of people do not have a problem with this, but a lot of people do.

She said that across diverse online communities the expectations are so varied that it is difficult to discuss them. It is equally difficult to know about those varied ethics and take them into consideration while still doing high quality AI/ML machine learning work? She thinks that the Twitter example about thread reader apps and who doesn't like them and who uses them anyway, highlights an interesting question - how does copyright law in particular and society in general think about who is a creator of content?

She provided an interesting example in the song "Folsom Prison Gangstaz". It is a mash up of Johnny Cash's "Folsom Prison Blues" with Eazy-E's "Luv 4 Dem Gangsta'z". She uses it when teaching and asks, does copyright exist to prevent this remix from happening without permission? Some students say "yes" or at least they say, copyright exists so that the creators can say yes or no. But the interesting thing in this discussion is that except for groups that are mostly people under the age of twenty-five, the only creator people spontaneously talk about in this discussion is Johnny Cash. They don't talk about Eazy-E's right to say "yes" or "no" to the remix. And the fact is that neither one of them had any input into whether the remix happened. She used this example to illustrate that the ideas about who is an author tend to center on certain types of creators and certain art forms. For example, there are artists who have made their reputation by copying other people's images and usually recontextualizing them in some way. Richard Prince [31] is well known for this, and when he gets sued, he tends to win. Jeff Koons [32] is another example.

While Kaufman talked primarily about copyright and the inputs to AI/ML. Sims brought up the issue of copyright with regards to AI/ML outputs. She showed an example of a machine-generated image and said that this is what happens when you let machines make photographs from inputs and AI creates. She asked who owns the copyright in this image? [33] She noted that under U.S. law, a work eligible for copyright protection must be an original work of authorship, which arguably some machine learning pictures can be taken for. But the U.S. also holds that in the U.S. copyrightable works must have a human author. The U.S. law Section 313.2 states the following - the office will not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author. Sims said that this opens a can of worms. What is creative input or intervention from a human author? If you train an algorithm how to create something from data sets, are you doing sufficient creative work to be considered the author [34].

In closing Sims said that there are no clear answers here, but there are people in various places around the world who are lobbying for new approaches that do create ownership rights regardless of human authorship. We need to wait and see.

6.3. Discoverability in an AI world

The speaker in this session was Andromeda Yelton, an adjunct faculty member at the San Jose State University School of Information [35] where she teaches courses in Artificial Intelligence (AI). Her objective was to discuss how libraries and other cultural heritage organizations are building and using AI and some of their concerns. She began her presentation by talking about four AI projects that supports traditional forms of discovery:

- *Annif*: https://annif.org, is a tool for automated subject indexing and is a project of the National Library of Finland. It allows you to upload the text for which you need index terms/subject headings, and it will do an analysis and provide some suggestions. It does need some human oversight, but she said that it it could radically accelerate a cataloger's productivity and allow them to catalog a much larger set of texts by giving them suggestions that they can evaluate, choose what works, and then just spend their time on the ones that are particularly challenging to catalog that a computer cannot handle.
- *Teenie Week of Play:* Another set of AI projects is the Teenie Week of Play [36], so-called because it was a week of investigating various computational approaches to the Charles Teenie Harris Archive [37]. He

was a famous photographer for *The Pittsburgh Courier.* The projects used AI to automatically shorten titles, extract locations and personal names, and look for the same person in different photos across the collection (*Note*: the archivist of this collection spoke in another session that I cover later).

- *Transkribus*: This is a comprehensive platform for the automated recognition, transcription, and searching of historical documents [38]. It's a project of READ-COOP, which is a European Union (EU) cooperative society organized for social benefit rather than for profit. It is open to non-EU members as well as EU members. It has been used in the Amsterdam City archives, the Finnish archives. It supports Arabic, English, old German, Polish, Hebrew, Bangla, and Dutch, and it can be trained to handle more languages and scripts. There is a demo on the READ-COOP website along with an option to download the software and install it to have a full set of features. Yelton said that as with the Teenie Collection, this can accelerate the human labor that is involved in archiving
- *Lasekompas*: This is a system that is used to recommend books to library patrons [39]. It is a project of the Danish Bibliographic Center, which is a public-private partnership that provides bibliographic data and Information Technology services to Danish libraries. It is a digitally-based dialogue that can support and emulate the traditional librarian-lender conversation with the aim to develop a better and more targeted recommendation of library content to users. It is also based on some novel metadata. The people driving the project worked with librarians in Denmark to have the librarians design a new vocabulary that better served their users' expectations. And today all new fiction in Denmark is cataloged with this new vocabulary. The vocabulary tries to convey the feelings that the reader might want to feel or the atmospheres that they might want the book to convey. (*Note:* If you go to the laesekompas.dk website, you will see various reader types that have adjectives next to them, such as fantastic, or mystic. If you click on any of those adjectives you can find other books that are cataloged with those adjectives.

Yelton went on to say that AI can also power novel forms of discovery. It can be used to create discovery systems that are unlike anything that currently exists in libraries or cultural heritage organizations. She gave an example of a project that she herself created entitled "Hamlet" [40] that is used to explore a database of graduate theses. The algorithm was trained on a corpus of about forty-three thousand Master's and PhD theses, mostly from science, technology, and engineering-type subjects. There was some metadata associated with it, but not much. It did not have subject access, nor did it have full-text search capabilities. While one could look at all the documents that came from the same department and hope that they had something in common, the fact is that there can be thousands of theses from the same department that do not always have much in common.

The theses with which she was working came from MIT and they have an Electrical Engineering and Computer Science Department, which has some five or six thousand theses in it. But she noted that some of the ones at the electrical engineering end of that department do not really have anything in common with the computer science theses. They're much closer to mechanical engineering or physics. And some of the theses from the computer science end of that department have nothing in common with electrical engineering - they are more math-oriented. Therefore, using departments for searching is not great for either colocation or discovery. She then used an algorithm called doc2vec [41], which trains a neural net, and it puts theses visually closer together or farther apart depending on how conceptually-similar they are to one another. It is successful only if the text of the theses had been digitized in a good way, but she ran into problems when the text came from old typewritten documents that were not correctly processed via Optical Character Recognition. In these cases, the Machine Learning algorithm struggled. She found that there was a lot of data in the database that isn't amenable to computational extraction. Some of it was originally written on a computer and submitted as a computer file but had been stored as a PDF or a print form of record.

She said that if you cannot search using visualization, you can use the "recommendation" method. For example, search for authors and search by titles, and this can be done on the Hamlet website right now. And you can find out what are the closest and most similar other theses to a given title or to the works of a given author. And if you know that one of them is relevant to your research, you can ask the neural net what else you might want to read. She gave other examples of searches that can be done by Hamlet or similar systems. For example, you can upload the first chapter of a work in progress, and it will alert you to similar dissertations.

Yelton then mentioned PixPlot [42], a project out of the Yale Digital Humanities Library that is used for exploring and visualizing a large corpus of manuscript images. They are using it for a body of about twenty-seven thousand images from the Beinecke Library. And like Hamlet, they are putting them closer together or farther apart depending on their level of similarity. Only in their case they are looking at images rather than text. She added that PixPlot has made this extremely large collection explorable and accessible by making it visual in a way that it might not have been done without AI (*Note:* I went to their web site and found it totally fascinating - definitely worth a look).

A third AI project that facilitates new methods of discovery is Citizen DJ [43]. This is a project undertaken by Brian Foo, a 2020 Innovator in residence at the U. S. Library of Congress (LOC). It allows anyone to freely make hip hop music using the free-to-use audio and video collections from the LOC and allows listeners to discover items in the Library's vast collections that they would never have known existed. Yelton noted this project not only makes the collection more accessible to people, but also has turned it into something with which people can make their own creative works.

She closed out her fascinating presentation by providing a list of challenges to the use of AI and she warned that the list is incomplete.

- First, data cleanliness as noted earlier when discussing her Hamlet project. If data is not digitized clearly, it will not be possible to feed it forward into Machine Learning discoverability systems in a way that makes any sense.
- Second, metadata is inconsistent. Institutions have their own best practices for how they create their metadata records, what fields they use, and even how they use some of those fields. The same institution can have different in-house style guides and best practices at different times. There may be some records that are extremely thorough and extremely accurate, and others that are extremely skeletal. *Remember* that computers need consistency.
- Third, typographical errors. There are lots of them in metadata records and full-texts and computers are not smart enough to know that they are the same word.
- Fourth, name changes. People change names all the time, serials publications change names, institutions change names. Therefore, when you have a data set that spans a long time period, figuring out how to match those up can be challenging.

Another challenge is data bias (Nancy Sims referred to this in her earlier presentation). Yelton said that she spends weeks on the issue of data bias in her classes so for the sake of time she would over-simplify the discussion here and noted that the data sets we use tend to be full of biases about gender, race, nationality - any number of things - and they may reflect stereotypes. Indeed, there are a lot of stereotypes, omissions, and over-representations in data sets. For example, facial recognition systems tend to be trained on white men and therefore tend correspondingly to do worse on women and worse on dark-skinned people, and especially badly on dark-skinned women. She mentioned a project, Gender Shades, that evaluates the accuracy of AI-powered gender classification systems [44]. She reinforced this issue by stating that this was one of the problems in the AI exploration of the Charles Teenie Harris photography archive that she mentioned earlier that covers Black people's lives in Pittsburgh in the middle of the last century. Also, the English language itself is wildly overrepresented in any sort of natural language processing. She noted that Google Translate works well for English, but that it works less well for minority languages. There are also problems that result from homogeneous AI teams where problems go unnoticed because they do not impact anyone on the team. Google deployed photo tagging without ever noticing that it did incredibly badly on Black people because there are not a lot of Black engineers at Google. It is easy to miss these problems when you do not have a diverse team (*Note:* At the end of his presentation noted earlier, this was Roy Kaufman's first question to consider when evaluating AI input).

Another challenge has to do with surveillance (think about Doctorow's opening keynote). The data that many AI systems are trained on is data about human behavior, behavior relating to the things that we click on online for advertising, face recognition cameras in public life, user behavior in libraries, etc. Bias in this instance can have a disproportionate effect on marginalized populations who are more likely to be surveilled already (*Note:* dana boyd covered this point with real-life examples in her 2020 NISO Plus talk that I mentioned earlier) [30].

Yelton's final challenge is the availability of the requisite resources. If an organization wants to develop artificial intelligence tools in-house, they will need software engineers who are comfortable with math. They will need money and the labor to assemble and label data sets. They will need time and cloud computing to train the AI models. As a result, building in-house artificial intelligence may be out of reach for a lot of cultural heritage organizations. While it is getting cheaper all the time, and partnerships between libraries, archives, museums, and other organizations, as was seen with the Teenie Harris Week of Play, do help. But it is not a realistic option for most cultural heritage organizations to do this in-house. Yelton believes that it is much more likely that most of the of high-profile AI tools will come from the vendor community, but her concern that scenario is the potential pricing. Will such tools only be available to highly-funded libraries? And even if the price points are accessible, will the tools require so much in-house knowledge to run, take care of, and maintain that only a small number of libraries can have them?

She closed by saying that she believes that there are a lot of opportunities offered by AI, but that there are also a lot of concerns. Hopefully, we as a community can have a discussion and move forward.

7. AI, metadata, and historical bias

Continuing the theme of the dangers of bias being drawn into to AI initiatives, this session had three speakers who shared their thoughts on the challenges of metadata creation for and by AI, and how standards and best practices can help address them.

7.1. Metadata creation

The first speaker was Dominique Luster, the Teenie Harris Archivist at the Carnegie Museum of Art, who spoke about how easy it is to instill bias if we look at things only from a single perspective. She sees historical bias in galleries, libraries, archives, and museums resulting from the idea of "white normativity". And this is the often unconscious and invisible ideas and practices that make whiteness appear neutral. Whiteness can be "seen" only when compared to something else such as Blackness.

She went on to say that the framework of archival metadata can typically be made of three components - content, context, and structure. And librarians are often led to believe or are taught in library school that archives are neutral. But she said that is not the case. Librarians are taught to use passive, neutral, or objective language as they create metadata records. But what if the language that librarians use is not truly passive or neutral and that they have simply been taught that it is? What if the language being used for archival metadata is within the context of white normativity and not within the context of the world and community that produced the material contained in the archive? Does the contextual relationship that is being framed match the actual real-life experience of the community represented through the collection being archived?

Luster echoed the warnings of prior speakers - Artificial Intelligence (AI) and Machine Learning (ML) outputs are only as strong as the data sets from which the algorithms are trained. If the data set(s) that are used to process the backlog collections are made by humans who are filling those data set(s) consciously or unconsciously with their own human biases, then everything that will emerge from the data will reflect those same cultural biases. She went on to that it is not simply about creating diverse or inclusive data sets. No, what she wanted to make clear is that these data sets and the work of libraries must be by design anti-racist and anti-oppressive. She also said that the use of machines does *not* speed up her work at all because the thought and the care that must go into this work takes a substantial amount of time. To create clean, clear, and culturally-competent, racially-conscious data sets for AI and ML projects takes time, is very complicated, and must be interrogated at every stage and at every layer, from whether you're using DACS [45] as an archival standard to a repository's processing scheme because machines can exponentially compound the impact of cultural inequities.

She closed with a quote from a poem by Amanda Gorman:

> "We seek harm to none and harmony for all.
> Let the globe, if nothing else, say this is true:
> That even as we grieved, we grew;
> that even as we hurt, we hoped;
> that even as we tired, we tried;
> that we'll forever be tied together, victorious,
> not because we will never again know defeat
> but because we will never again sow division [46]".

7.2. *Return on investment: Reframing AI and metadata*

The second speaker in this session was Michelle Urberg, Affiliate Associate, Maverick Publishing Specialists, who opened by saying that she would be speaking about the effects of naming/subject headings/metadata within the context of historical bias and how that bias can hinder the ability of Artificial Intelligence to leverage metadata to its fullest in a responsible and equitable way. She added that she is not only talking about financial bottom lines, but also about social responsibility, particularly in the archival library and academic publishing space. Urberg said that she believes that Dominique Luster, the first speaker in the session, did a wonderful job of providing a theoretical context, and that she (Urberg) purposely wanted to sit alongside of the theoretical and show some specific things about the power of naming and why librarians must be careful. She then went on to provide several concrete examples of how historical bias has caused problems and I will not repeat them here as she does it so much more eloquently in her own words in a separate article that appears elsewhere in this issue (definitely worth a read).

Throughout, her message was loud and clear, and she closed by saying that naming and metadata is everything and that historical bias in metadata creation suffers from language decay. It is a more insidious problem than refusing to remove outdated terms or naming conventions in controlled vocabularies. It is a statement of cultural values and an unconscious bias inherent in cataloging and metadata creation. To train artificial intelligence requires constant vigilance to train a system to behave equitably and to respond to the changing needs for user discovery. She said that we as a community face a challenge here today and need to begin a discussion about transformative naming with metadata as a strategy to improve AI technology for discovery.

7.3. The publisher's perspective

The final speaker in the session was Joris van Rossum, Director of Research Integrity, International STM Association. A past speaker at several NFAIS Annual conferences, Joris is very experienced in scientific scholarly publishing, and he opened his presentation with a broad view of science in general and how it has evolved throughout the ages.

He noted that science started as being quite observational, looking at the stars, coming up with theories, and gradually evolving into a more empirical and experimental science in the 17th and 18th century. The invention of the computer changed science fundamentally, taking it from theoretical to computational. He said that today we are in an era of data science and with the amount of data being available, the practice of science is also changing. With vast amounts of data within reach, with an increase in computational power, and with the introduction of new technologies, science will soon become "smart science". He noted that AI is one new technology that is very promising and that it has the potential to fundamentally change and improve the practice of science. AI can basically test a hypothesis against vast amounts of data. We are now able to plow through enormous data sets, come up with new hypotheses, design new theories, and explore new connections that a human could never be able to make on their. It can also do research. And yes, it can even run entire labs and write research articles. There is already some experimentation with writing books using AI (*Note:* if you want to learn about AI-run labs take look at Arctoris [47]. I heard them speak at the World Chemistry Congress in August 2021 and I was blown away!).

With all of this, there is a lot of potential to change science, to improve science, and to improve the impact that science has on society. He said that when he thinks more deeply about AI there are several components in it. First, the input data, that trains the AI systems; second, the algorithms - the computer code that does the AI; third, the outputs of the AI process; and fourth, the use and the applications of the AI outputs. And academic publishers are involved in all four components. First, publishers are providers of high-quality data. In that component, we play an important role. Second, publishers are developing AI tools, we use AI tools, and we communicate the AI outputs through our journals. He noted that STM is preparing a white paper on AI, science, and ethics which they plan to publish by the end of April 2021 [48]. In preparation for the White Paper, they did a survey that asked publishers about how they are currently using AI?

To their surprise, they learned that publishers are already using AI in diverse ways. It is used to recommend articles to readers, as Amazon does, by learning from what readers consume and feeding them new content. This has been happening for a few years now. AI is also used to identify the right journal for the submission of manuscripts and to find the right editors or reviewers in the process. AI is used to identify the quality of English for submitted manuscripts and to determine the appropriate workflow for a particular manuscript. The even found examples of where AI takes content and automatically write books [49]. He added that what is especially interesting on the horizon is that AI allows publishers to detect and prevent fraud, which is an increasing issue for them. If you think about the ability of unethical researchers to duplicate submissions, to plagiarize, to manipulate data and images, etc., being able to apply AI to prevent those activities will improve the integrity of published science and increase trust in scientific outcomes.

He then went on to identify the downsides and risks to using AI and said that there are three big risks. First, as noted by previous speakers on this topic, AI outputs are only as good as the data that feeds the algorithms. Using flawed or wrong data will ensure that whatever comes out is worthless. So that is an important consideration, making sure that the data that is used as input is the right data, correct data, and well-selected data. Second, if the models are flawed AI will not work well and can possibly do harm. This risk is enhanced by the fact that AI is an opaque and difficult technology. People do not easily understand the algorithms, and sometimes the companies using them do not understand the algorithms themselves. The third risk is that by its very nature AI teaches computers by identifying patterns in existing data and existing processes. It amplifies the past and the present and depending on the data that is used, AI predictions can lead to discrimination. Historical bias can lead to discrimination and that is something that we must work hard to prevent. But van Rossum believes that for science there is another risk related to historical bias in that AI can potentially reinforce contemporary paradigms and structures, i.e., what has been successful in the past will be used to predict the future. But real science, as Thomas Kuhn, the famous historian or the philosopher of science has said, all significant breakthrough comes by *breaking* patterns, by breaking paradigms, and by introducing new ways of thinking. The risk of AI in science is that by strengthening existing patterns and paradigms, it can suppress innovation and ensure that we never get out of existing paradigms. This, he believes is the real threat that AI poses to science and therefore to society, especially when AI is used in the evaluation of science and in peer review. He said that we must be very careful not to allow this bias to impact science in a negative way and ultimately suppress innovation and break-throughs.

He went on to say that the use of AI has risks, not only for science, but also for any discipline or business in which it is used. There are organizations and governments working on ethics principles. Not only STM, but also the Organization for Economic Co-operation and Development (OECD), the European Union, the U.S. government, and others. If you look at different sets of guidelines, they have certain general principles in common - such as AI should benefit society; it should respect the rule of law and privacy; it should be robust, safe, secure, and accountable; it must be transparent and it must have human oversight.

His final remarks were related to the importance of the metadata that is associated with the raw data used by computers in AI processing and how that metadata is created. First there is the scientific content in the journals and the metadata here is created by the author via keywords, bibliographic information, etc. Then additional data is added via the Abstracting and Indexing Services. And of course, there are the datasets associated with an article that may be fed to a repository where additional metadata may be created. Is this enough? And do we need to have more processes to create the right metadata, ensuring the appropriate use of that data for AI?

His closing remarks were that he believes that AI has the potential to support and revolutionize science and therefore help society even more than it does today. It plays an important role for publishers as well, since they are involved in a lot of the AI components. However, there are risks, so we need to work on ethical principles. And again, the production, creation, dissemination of high-quality data and metadata is crucial, and therefore is an ongoing focus for all scholarly scientific publishers.

Remember that an article based upon Michelle Urberg's presentation in this session appears elsewhere in this issue of *Information Services and Use*. As an FYI, anything she writes is worth a read. Based upon her presentation at the NISO Plus 2020, she wrote a paper on Digital Humanities and Standards [50]. It was in part about digital humanities projects and the metadata that supports them internally and externally. But it was also a call for project creators, publishers, aggregators, and professionals working on metadata and standards to start a conversation about how to incorporate digital humanities projects into the scholarly communications lifecycle in spaces where books and journal articles have dominated for decades. I very much enjoyed it!

8. Privacy: Global perspectives

Following the session above there were two parallel sessions at 9:30pm EDT (sessions were held from 10:00am - 3:00pm EDT and again 6:30pm - 10:45pm EDT each day to accommodate different time zones). This session on privacy had three speakers, two of whom have submitted papers based upon their presentations that appear elsewhere in this issue of *Information Services and Use*.

8.1. Thinking with the General Data Protection Regulation (GDPR)

The first speaker was Andrew Cormack, Chief Regulatory Adviser, Jisc [51], who spoke about the General Data Protection Regulation (GDPR). I must give him credit - he was giving his presentation at 2:00am UK time! He opened by saying that one thing everyone knows about Europe is that the EU has a very strong privacy law - the General Data Protection Regulation or GDPR [52]. The goal of his talk was to get people to view the regulation not just as a law that ties your hands when doing outreach activities, but rather as a useful way to think about designing systems and processes.

Cormack focused on the much less referenced rules that relate to the free movement of personal data. The GDPR is explicit in its very first article about helping the movement and use of personal data if it is done in a way that is safe for individuals. He talked about three myths and the first was that GDPR is not primarily about individuals. For although the GDPR is al; about protecting individuals, its duties fall almost entirely on the organizations - which the Regulation calls "Data Controllers - that decide why and how to collect, store, use, share and dispose of (or anything else within the broad definition of "processing") data. These entities must ensure compliance with seven Principles, set out in Article 5, which are key requirements for all system designers. All the GDPR principles are aimed at those organizations. And those principles are a useful guide to designing safe products, services, and other activities. For example, accountability requires not only that organizations are compliant, but also that they can demonstrate how that they are compliant. Cormack said that *before* an organization starts to use personal data, they *must* think about (1) the design of their systems and processes, (2) safeguards against error and misuse, (3) how they will operate them safely, and (4) how they will ensure that those plans happen. The key point is that the focus here must be on the individuals and groups whose data the organization processes, and not on the organization. And the GDPR provides a tool - the Data Protection Impact Assessment (DPIA) to guide that thinking.

DPIA is mandatory for large-scale and otherwise high-risk processing. But it is a useful tool for thinking about smaller activities as well. And once an organization has done a DPIA he recommends that they publish it to demonstrate to their users and other stakeholders that the organization is taking care of their interests.

He went on to say that another principle both of law and design is purpose limitation. This requires that the organization think clearly and precisely about *why* they are collecting, processing, and using personal data. Multiple purposes may be okay, but the organization must be clear in their own mind and in their documentation what those purposes are. “In case it comes in useful”, is not a convincing purpose either for regulators or for stakeholders. Once having set out a purpose or purposes, the organization must avoid creep beyond them. They must ensure that it has a lawful basis for that purpose. Is it something that the organization needs to do to fulfill an agreement with the individual? For example, to pay a salary or deliver a service that they requested. Or something that an organization is required to do by law, e.g., telling a tax office about the salary? or something that is needed to save a life or prevent serious injury. Or something that is in the public interest and where an organization is best placed to do it. Or something that is in the interest of the organization, of individuals, or third parties it may work with. Each of these has its own conditions that the design must satisfy. For public interest and other legitimate interests, an organization must balance their interests with those of the individuals whose data they intend to process. He noted that if it is difficult to meet those conditions, then an organization needs to rethink either their design or whether they should be doing this at all.

Cormack said that the second myth is that the GDPR is about preventing process. That is not the case. The GDPR is about allowing processing that is necessary. The term “necessary” has a very specific meaning, i.e., that there is no less intrusive way to achieve the purpose. It forces the organization to think again about good design practice - about minimization. How little data does the purpose need? How little processing? How little disclosure both internally and externally? And how soon can the organization dispose of it? GDPR and its guidance recognize lots of technologies as contributing to this.

The final myth that he put forth is that the GDPR is mostly about choice or consent. He said that this also is not true - the GDPR is about notice. With very few exceptions, people must be told the natural consequences of the situation that they are in or are about to enter. Most of what an organization must tell them is the product of the thinking in the first two stages: Who is processing their data? What processing is being done, including the legal basis. Why, including the purposes? How long this will continue, and what happens to the data when it’s stopped. Who else is involved and where are the located? And how to exercise their rights over their data. Sometimes, but far less often than claimed, individuals will have a free choice whether to give an organization their data. But remember the five legal bases. If the organization is offering them a service, or required by law to process the data, or saving a life, or serving a public interest, then their choice isn’t free. He said that he believes that true consent will be appropriate when an organization wants an individual to volunteer information to get into a deeper relationship with them, not to discover whether they want a relationship at all. If an organization cannot find a basis for that initial relationship among the first five bases, they should rethink their plans. He believes that thinking system design using the GDPR helps an organization to meet the expectations of their users, customers, and wider stakeholders. He said that the advantages of GDPR thinking - to get more benefit from data, while managing the risks of its use.

Cormack submitted a paper based upon his presentation and it appears elsewhere in this issue of *Information Services and Use.* It contains a lot more detail and I highly recommend that you read it. Cormack made me look at the GDPR in a very positive way. He convinced me that the GDPR is a valuable tool, not a major limiting regulation, but rather a rich source of guidance for system and process designers.

8.2. Health wearable and apps: A changing privacy landscape

The second speaker was Christine Suver, Director of Research Governance and Ethics at Sage Bionetworks. The focus of Suver’s talk was digital health - particularly regarding mHealth, i.e., the use of consumer wearables and mobile applications to collect health data. She said that mHealth is a growing field and it is estimated that the global number of health app is going to reach about 100 billion U.S. dollar by 2023. It is growing because of new technology that facilitates the collection of data that provides a much broader and complete view of a person's health. The technology can continuously monitor some aspects of a person's health, e.g., like how much they exercise, their heart rate, glucose level, blood pressure, etc. It is a rich data collection that occurs almost automatically through sensors. And that rich data set can supplement data that is collected when a person visits their doctor. But what privacy rules apply to the mHealth domain? Suver noted that many countries have enacted some sort of privacy regulation and that a good example is the GDPR that was discussed in the prior presentation by Andrew Cormack. She added that the GDPR is one of the most important pieces of legislation, as it's applied to twenty-eight different countries in the European Union (EU) and three additional ones that are not part of the EU.

She went on to say that in the U.S. there is no comprehensive regulation on data privacy. Data privacy is handled by a specific domain. For example, there is regulation about data on communication, regulation about data on finance, and regulation on health-related data. Some states have started to enact data privacy laws, such as the California Consumer Privacy Act [53] enacted in 2018 and the New York SHIELD Act [54] enacted in 2019. Some other states are considering different privacy laws. But most of the privacy laws in the U.S. are not directly controlling how health data needs to be regulated. In the U.S., the landmark regulation for how to protect health information is the Health Insurance Portability and Accountability act of 1996 [55] (HIPAA). Suver said that even though individual countries have developed their own privacy regulations, or in the U.S., different states are considering their own privacy regulations, there are some global privacy principles that seem to be universally accepted. Examples of these principles are: (1) that the collection, use, and processing of personal data should be at minimum lawful, purposeful, transparent, and limited in time and scope; and (2) the data must be secure and used for an organization's intelligence only. These privacy principles and the privacy regulation apply to processing, collection, and use of personal data

Suver noted that there are different types of personal data. Health information is a special category of personal data. It is more sensitive and requires much more protection. But what about mHealth data? What about health-related information that is collected continuously through sensors? Is that considered medical data and is it a special category of data? She said that in the U.S, the answer is that "it depends". In the U.S., apps that monitor health and lifestyle are not regulated under HIPAA unless they are used in a regulated arch context, or they are used to make health-related decisions It means that the Food and Drug Administration (FDA) does not care about controlling lifestyle apps that track a person's diet, exercise, and sleep even if the apps or wearables are targeting children. Yet, these apps are handling personal and sensitive information. Yes, the data collection is lawful and legitimate based upon the consent that people provide through accepting terms of services or privacy policies. But even if the data is collected anonymously, if it includes geolocation, for example, it can show a pattern of interaction between an individual and the device. Therefore, even if individuals are not identified, the data gathered from those apps and wearable devices can be significant and provide a lot of information. Indeed, an app that records a person's location may pose a greater risk to privacy. She noted that during COVID, there has been an explosion of contact tracing apps and in a study of about five hundred COVID-related apps it was found that some of these are collecting information that is not necessarily needed for contact tracing, such as data from phones and phone cameras.

Suver added that there are some situations in which a person may not want anonymity, especially in the case of personalized medicine. Health care professionals want to be able to collect information about an individual so that they can provide tailored medical services. And for that, they want to collect information such as a person's medical records, genomic information, and information about their environment, and lifestyle. She said that there is not a one-size-fits-all policy when considering privacy law. But one of the principles that is important is to obtain the data by lawful and fair means and, when appropriate, to be able to obtain consent. Usually, consent is obtained through terms of services and privacy policies. Yet most people, herself included, never read a complete s of service. They just click and "accept".

In closing, Suver asked why suppliers do not use the concept of "continuous consent [56]" instead of asking people to read a privacy policy or to agree to terms of service at the time an individual downloads an app or uses a wearable device and noted that she would love to discuss this concept (sounds interesting to me and I would love to know the legal implications).

Suver also submitted a paper based upon her presentation and it appears elsewhere in this issue of *Information Services and Use.*

8.3. Personal information protection law

The final speaker in this session was Judy Bai, Director of Business Development at Digital Science who spoke briefly about China's Personal Information Protection Law [57] (PIPL) which will be enforced as of November 1, 2021. This was news to me and if it is to you as well, I suggest that you to take close look at this new law coming out of China - especially if you are dealing with customers in that country.

Bai said that since her organization is just starting to establish a presence in China, they knew that it was important to be aware of the China legislation and policy development, to review her organization's existing policies, and make necessary preparations where needed. She said that she wanted to share some of the preliminary information that they have gathered on data protection laws in China.

With measures to ensure privacy becoming increasingly prioritized worldwide, many countries have framed relevant laws and regulations on personal information protection. The most notable being the European legislation, GDPR, which Andrew Cormack discussed at the beginning of this session. On October 21st, 2020, China released its draft Personal Information Protection Law (PIPL) for a months-long public consultation after the first review by the Standing Committee of the National People's Congress.

Before 2020, China had already implemented a series of laws and regulations that covered the protection of personal information. For example, China's Cybersecurity Law [58], which came into force in 2017, governs the protection of personal information with a focus on the protection of information in the cyberspace, the protection of critical information infrastructure, and the regulation of network operators. And in July 2020, a draft of the China Data Security Law [59] was also released for public comment. But the draft PIPL marks China's first attempt to systematically and legislatively define, establish, and integrate the provisions on the protection and regulation of personal information. It is regarded as a major milestone in China's legislative effort to establish a set of comprehensive regulations around data privacy. The draft PIPL is a concise document under eight thousand characters. And it comprises a set of eight chapters with seventy articles. And those familiar with GDPR will find some similarities in the draft PIPL when reading it for the first time. She added that, indeed, some concepts are inspired by the GDPR.

Among other things, the draft PIPL sets out data protection principles; specific rules for the processing of both personal information and sensitive personal information; the rights of individual data subjects; and penalties for breaches. She mentioned a few key features of the draft PIPL which are worth highlighting. One is extraterritorial application - in general, PRC laws do not have extraterritorial effect. However, the draft PIPL appears to follow the approach taken by the GDPR and will have a long-arm extraterritorial application to any personal information processing activities of organizations carried out beyond China's geographical borders. *It is also worth noting that a non-PRC established organization that is subject to the PIPL due to this application should appoint a representative within the country to deal with data protection-related matters.*

The second feature is a new legal basis for data processing. Under existing China laws and regulations, a data subject's consent has been established as the only legal basis of processing of personal information in China. In the draft PIPL, new legal bases are introduced for personal information processing depending on why the processing is necessary. For example, for legal duties, or obligations, or to respond to a public health emergency, or to protect the life, health, and property of a person in an emergency.

Also, the issue of data localization and cross-border data transfer has been the subject of much discussion and debate since the Cybersecurity Law came into force in 2017. Under the new draft PIPL, there are positive developments providing more alternatives for international companies to manage their cross-border data transfers in a legally-compliant manner and to some extent like the thinking behind binding corporate rules under the GDPR.

And finally, hefty fines are being applied. She said that serious legal consequences have been historically absent from Chinese data protection laws and that the draft PIPL takes a totally different approach. Organizations violating the law could be imposed with fines up to fifty million Chinese yuan. That is equivalent to 7.5 million U.S. dollars or 5% of an organization's prior last year's annual turnover together with business suspension, license revocation, and potential civil or criminal liability.

She said that with this information in mind, her organization is now ready to proceed to answer some questions important to the company. First, will our company be regulated by China's PIPL? One common misunderstanding of the PIPL is that it is only applicable to internet firms, such as these tech giants, Tencent, Baidu, etc. But we learned just now that if you have a business running in China, you will be regulated by PIPL, as there is always personal information, such as email address and phone numbers that gets collected and processed during business operations and interaction with customers. Even if your company does not have a physical existence in China, it may still be regulated if your company processes the personal information of the people in China for the purpose of providing products or services to people in China or analyzing and evaluating the activities of the people in China. Examples include selling products via online shops to Chinese consumers, providing online language-training courses, or using AI-based technology to surveil people in China, such as facial recognition, location tracking, profiling, etc. She said that they also just learned that under the PIPL entities outside of China that collect and analyze data for these purposes will need to appoint a data protection representative or organization within China to manage these matters.

How does this impact an organization's Information Technology infrastructure and applications? She said that to answer this question the following information is required: (1) whether the personal information processed by a company can be transferred out of China. We know that under PIPL that the cross-border transfer of personal data to foreign authorities can be achieved, but it will still require prior approval from Chinese regulators; (2) whether any sensitive personal information is being processed. The GDPR sets out an exhaustive list of special categories of personal data. The PIPL list of sensitive personal information is shorter, but it can cover a broader scope of personal information when compared to GDPR depending on how strictly or loosely this definition of sensitive personal information is interpreted; (3) what data classification and retention techniques are deployed in an organization. This will require a company to deploy relevant techniques to classify information and implement a proper data retention policy to delete relevant information that is no longer needed for the original purpose of collection; and (4) is the company using any mobile apps to communicate with people or deliver services to China. The Chinese Government has launched several campaigns in the past few years on mobile apps to combat the illegal collection of personal information. Many apps have been required to change or have been removed from the mobile apps store. Therefore, if an organization uses mobile apps to communicate with people or deliver service to clients in China, they need to pay attention in the app development stage and ensure that the process of the permission access request is proper.

She said that the common measures that can be taken to protect personal information that meet the compliance requirement imposed by the draft PIPL can be divided into two categories - technical measures and organizational measures. These are as follows:

- *Technical measures:* General security control measures; encryption; de-identification measures; data classification/date retention/data loss prevention (DLP) measures; 'privacy by design' and 'privacy by default'.
- *Organizational measures*: Running a data protection impact assessment; staffing; training.

She did not go through the list one by one, but said that those in the audience should bring these measures to the attention of their organization's technical team as well as legal colleagues for assessment and planning.

She stated that as China's first comprehensive law in personal information protection, the PIPL strengthens the protection of personal information while considering the complexity of economic and social life. This draft drew intensive media and public interest from legal professionals, academics, and business representatives. Many studies were conducted to compare the draft law with the GDPR and other data laws around the world. The conclusion is that it appears that the gap between PRC regulation and the GDPR of the European Union is closing regarding personal information. The PRC regulation aims at covering the entire cybersecurity area, but the GDPR appears to still be more comprehensive, especially regarding accountability, regarding the distinction between data controller and data processor, etc.

In closing, Bai said that given the potentially wide application of the PIPL and the measures necessary for compliance with Chinese law, companies expecting to run business or to be governed by this law should immediately begin to monitor developments and review their policies and practices in preparation for the November 2021 enactment of this significant new law. And they should also factor in the costs that may be incurred to ensure personal information protection while planning their budgets for the near future.

9. Miles Conrad lecture

A significant highlight of the former NFAIS Annual Conference was the Miles Conrad Memorial Lecture, named in honor of one of the key individuals responsible for the founding of NFAIS, G. Miles Conrad (1911-1964). His leadership contributions to the information community were such that, following his death in 1964, the NFAIS Board of Directors determined that an annual lecture series named in his honor would be central to the annual conference program. It was NFAIS' highest award, and the list of Awardees reads like the Who's Who of the Information community [60].

When NISO and NFAIS became a single organization in June 2019, it was agreed that the tradition of the Miles Conrad Award and Lecture would continue. The first award was given ton James G. Neal, University Librarian Emeritus, Columbia University. This year's award was presented to Heather Joseph, who has served as the Executive Director of the Scholarly Publishing and Academic Resources Coalition (SPARC) since 2005. In that capacity, she works to support broadening access to the results of scholarly research through enabling open access publishing, archiving, and policies on a local, national, and international level. Joseph is also the convener of the Alliance for Taxpayer Access, a coalition of universities, libraries, patient advocacy groups, consumer groups, and student organizations who work to ensure that the results of publicly-funded research are openly-accessible to the public. The group has been a leading voice on U.S. open access policies, including the landmark public access policy issued by the National Institutes of Health (NIH), and the recent White House Directive mandating public access to publicly-funded research across all U.S. science agencies.

Prior to coming to SPARC, Joseph spent fifteen years as a publisher in both commercial and not-for-profit publishing organizations. She served as the publishing director at the American Society for Cell Biology, which became the first journal to commit its full content to the National Institutes of Health's pioneering open repository, PubMed Central, and she subsequently served on the National Advisory Committee for the project. Joseph serves on the Board of Directors of numerous not-for-profit organizations, including the Public Library of Science. She is a frequent speaker and writer on scholarly communications in general, and on open access in particular.

In her presentation, Joseph provided a brief look at how scholarly communication has evolved through the lens of her diverse career and she reinforced the importance of having a good leader and a trusting, strong mentor early in one's career. But the focus of her presentation was the Open Access movement - how it was born, how it has evolved, where it is now, and what still needs to be done to ensure that everyone in the word can share their knowledge and have access to the knowledge of others. She noted that to be truly effective, the strategies/solutions for improving the knowledge sharing system also must comprehensively address *all* the primary social justice principles: access, participation equity, and rights. She said that achieving these goals is what drives her, but added that she did not start out her career with this in mind. It has grown and come into focus over the arc of her career. She added that she was extremely fortunate to have had the chance to work for and with some incredibly smart, visionary, and generous people who gave her incredible foundational opportunities from the day she started working in this arena.

Joseph has been a pioneer in Open Access and has been involved every step of the way. While her presentation was both informative and enjoyable, the article based upon the presentation covers far more detail on the evolution of OA and it appears elsewhere in the issue of *Information Services and Use.*

10. Metadata and discovery

This session opened with a look at the Open Discovery Initiative (ODI) presented by Geoff Morse, Interim Head of Research Services, Northwestern University Libraries, and Ken Barnum, Senior Program Manager, University of Michigan. The goals of this initiative are to: (1) define ways for libraries to assess the level of content provider participation and for discovery services to affirm how they use that content; (2) help streamline the process by which content providers work with discovery service vendors; (3) define models for "fair" linking from discovery services to publishers' content; and (4) determine what usage statistics should be collected for libraries and for content providers. The initiative goes back about a decade when it was first proposed at an American Library Association Annual meeting in 2001. The first recommended practice was released in 2014. Three years later, in 2017, a revision process was started and last year, on June 24, 2020, an updated recommendation was released [61].

Morse said that the value proposition for the stakeholders is as follows:

- *Libraries/Users*: Finding relevant content is simpler when it's all in one platform. ODI makes it easier to understand which resources are included in discovery services.
- *Content Providers*: Participation in discovery services makes content more valuable and discoverable, increasing usage and decreasing the likelihood of cancellations.

- *Discovery Providers*: Participation in ODI increases transparency, improving customer satisfaction and retention.

He noted that all three groups are part of the ODI Process. Each group needs to provide input and has an impact on the other, and that communication needs to happen between and among each of the groups.

The role of the Content Provider is to: (1) provide high-quality metadata; (2) fair linking which is essential so that libraries can choose the platform to which they want to link; and (3) open access so that whenever content is open, the content will use the "free-to-read" metadata standard to let users know that the content is OA. There is a complete checklist to which content providers comply.

The role of the Discovery Provider is to: (1) ensure the transparency of what is included in the Discovery System; (2) ensure that high-quality metadata is made available to both libraries and content providers at both the collection level and the title level; (3) ensure that collection level metadata is provided in downloadable form; and (4) to provide fair linking, metrics, and an indication of open access content when relevant.

The role of the Librarian is to: (1) ensure that the discovery provider's configuration guidelines have been followed; (2) document all configuration decision; (3) assign staff to oversee specific areas of configuration; (4) confirm that subscribed content is enabled in discovery; (5) develop training to meet different users' needs (including library staff!); (6) review all system upgrades even if they are performed by the vendor; (7) complete and publish a library conformance statement; (8) follow up with vendor partners on their conformance; (9) advocate increasing ODI conformance for Content Providers and Discovery Service Providers.

In closing they provided reference materials for the stakeholders which I include below:

Resources for Content and Discovery Providers:

- Content Provider FAQ: https://www.niso.org/standards-committees/odi/content-provider-faq
- Implementation Guide: https://bit.ly/2Wblk7W
- Conformance Checklist Templates & Statements (note that the goal is transparency, not perfection): https://www.niso.org/standards-committees/odi/conformance
https://www.niso.org/standards-committees/odi/completed-statements

Resources for Librarians:

- FAQ and Talking Points: https://www.niso.org/standards-committees/odi/library-talking-points
- Publishers Discovery Configuration Guides: https://niso.org/standards-committees/odi/configuring-content-providers
- Conformance Statement Checklist: https://groups.niso.org/apps/group_public/download.php/24607/ODI%20Conformance%200Checklist%20Template_Library_2020.docx

General Resources:
Website: https://www.niso.org/standards-committees/odi
Mailing List: http://groups.niso.org/lists/opendiscovery/
ODI Updates: https://www.niso.org/standards-committees/odi/updates
Twitter: https://twitter.com/NISO_ODI

Note that a detailed paper on this initiative by Morse and Varnum appears elsewhere in this issue of *Information services and Use*.

11. Open Access analytics

Sara Rouhi, Director of Strategic partnerships at PLOS, introduced a presentation given by Tim Lloyd, CEO of LibLynx, on a case study that they are both working on the topic of next-generation Open Access (OA) Analytics. Sara gave a brief overview on the background of their work to provide context. She said that this study relates to the new challenges that publishers and librarians currently face in assessing the impact of traditional publishing agreements versus read-and-publish agreements. With the latter agreements, a large component of the cost is to cover the expenses related to cover the cost of publishing of content that will, as a result, be Open Access. The historic metrics that COUNTER [62] provides that are incredibly useful for assessing paywalled articles and content no longer really apply when OA content can be available anywhere at any time around the globe, with no need for IP authentication, which is obviously the primary means by which COUNTER looks at usage statistics. She said that the PLOS-LibLynx partnership is attempting to determine what the next generation of usage statistics will look like, and how those statistics will inform how librarians and other stakeholders look at the impact of an agreement with PLOS or any other OA publisher in determining whether to renew these agreements? And at a more general level, what can these metrics tell us about the impact of the research that is contained in an OA paper?

She then turned the podium over to Tim Lloyd who said that the project is in the experimentation stage and that it has three goals: (1) to understand stakeholder needs from open access analytics; (2) to provide COUNTER Reports to meet the community's immediate needs to better understand the impact of open access content published by PLOS; and (3) to develop next generation analytics that can meet the needs of the more diverse use cases that they are seeing in the open access environment. He said that the stakeholders in the overall process are the institutions that are generating the published research; the publishers who publish open access content; the authors who write up the research; the community that is interested in reading about the research; funders who pay for the research; and there are the many diverse intermediaries who perform a variety of functions that support the publishing workflow. It is a complex landscape with each group looking at metrics from a different perspective and there is now a demand for a broader range of metrics beyond the traditional analytics provided by COUNTER. For example, metrics that answer granular questions such as "which organizations access what OA content when and from where?" Lloyd said that they have a lot of other attributes with regards to the reporting of metrics. For example, in use cases where the usability of the data is more important than the scalability, it is not hard to imagine that visually-rich layouts that make for easy consumption would be of interest. There is also immediacy. COUNTER reports are monthly and there could be other calendar-based formats. Or a specific use case may need real-time access to analytics.

He went on to show some sample reports and closed by saying that he would like to put forth three questions to the audience as a foundation for further discussion: (1) "How has your thinking about usage data and other volumetrics changed, given the accelerated push for open access in the last eighteen months?" (Lloyd and Rouhi are seeing an acceleration in interest for having conversations around this area); (2) "Where can these next generation metrics support your thinking about 'impact' from the researcher/faculty or library/collections perspective?" and (3) "How can these new metrics work alongside existing COUNTER metrics to present a broader understanding of the value of open access publishing?

He asked if anyone has feedback to contact both him and Sara. Note that Lloyd provided a manuscript based upon his presentation and it appears elsewhere in this issue of *Information Services and Use*. It provides some samples of their reports.

12. Misinformation and truth: From fake news to retractions to reprints

This was a good session, but I am going to keep my comments brief only because all those involved in the session submitted a joint paper that discusses the topics that each speaker covered. There is no sense in my repeating their words. I will just give a summary and recommend that you read their article plus a related one from the 2020 NISO Plus Conference that caught my interest last year on the preservation of TV News because of biased distortion of the news [63] for political purposes (a somewhat topical subject!).

12.1. Scientific fact checking

The first speaker was Sylvain Massip, Co-Founder of Opscidia [64]. It is a French start-up that attempts to promote the re-usability of research outside of academia. He said that there are three reasons that Society as a whole does not use research results: (1) most of the material is behind a paywall, hence an awareness of the content and access to it is difficult; (2) reproducibility - if you have only one article stating something, it is difficult if you are not an expert in the field, to know whether the conclusion of the article can be trusted; and (3) discoverability - with around two million articles published every year, it is difficult to find the information that you need. The purpose of Opscidia is to remove these roadblocks to ease the reusability of scientific results by society as a whole. He said that they have an open access publishing platform that is Diamond Open Access [65] and therefore free to authors and readers alike. They fund the full enterprise with other services that are scientific text analysis tools and services.

The main focus of Massip's talk was his organization's work on scientific fact checking. It is a project that has been running for about a year and which is funded by the Vietsch Foundation [66]. His company built a prototype that is used as follows. The user enters an input statement, such as "does Agent X cure, cause, or prevent Disease Y?" Opscidia then selects the right corpus from Europe PubMed Central and from that body of articles they develop three indicators to say whether specific articles back or contradict the claim. The main objective of this prototype is to demonstrate that they can use open access scientific articles to help people make sense of different claims that they may have read about. Massip went on to describe the indicators (they are detailed in the full paper based on this session) and concluded by saying that they have built a pipeline based on three indicators to try and detect scientific consensus and to help the public understand what a research article is, what scientific consensus is, and how it all works. They want to help people discover by themselves that one must not simply accept the results of one study. They also believe that it is very important to demonstrate that open access has applications outside of academia, and that OA can be useful to fight fake news.

12.2. Retracted articles

Randy Townsend, Director of Publishing Operations at the American Geophysical Union (AGU) said that at AGU they are paying more attention to geographic boundaries. Territories that are the cause of disputes between two or more countries are being captured in AGU maps, phrased in a way that could lend to a case where one of the countries could lay claim to that region. AGU follows the Unite Nation's guidelines for the naming of territories, but AGU continues to see inappropriate names in manuscript submissions. He showed a figure that came from a paper containing China's unsubstantiated nine-dash line claim of territory over the South China Sea. According to this map, all the islands that are in dispute between China and other countries in this region, such as Brunei, Indonesia, Malaysia, Singapore, Thailand, and Vietnam, belong to China. The nine-dash line was not in the original submission, but was added sometime during peer review after the co-authors had already reviewed the submission. When it was brought to AGU's attention, the authors were offered the option of correcting the article by replacing the figure or retracting the entire article. Townsend had discussions with the authors and saw that there was clearly pressure coming from the Chinese scientific community to retain that nine-dash line. If the authors did not come to an agreement, AGU would have retracted the article so that they could remain unbiased in the political debate. Ultimately, in this case, the paper was corrected, but there was clearly an impasse among the authors.

12.3.

Caitlin Bakker, Research Services Librarian, University of Minnesota, continued the theme of retracted articles. She said that current processes, practices, and systems do not fully correct the scholarly record when an article is retracted, and this is known because retracted publications continue to be used. She went to say that one of the cornerstones of evidence-based medicine is the concept that one would use the best available evidence when making health care decisions, but doctors generally spend less than three minutes seeking out that information. While it is impossible to find all of the original research on a particular question in three minutes, it is possible to find a systematic review. Such reviews are the "go-to resources" for doctors, and they are taught and encouraged to use these resources in many medical schools and programs worldwide

Bakker said that she is currently working on a project with colleagues Sarah Jane Brown and Nicole Theis-Mahon where they are looking at retracted publications in systematic reviews, particularly in the pharmaceutical literature. And they found that in a sample of about fourteen hundred retracted papers, two hundred and eighty-three (20%) were cited over a thousand times in systematic reviews. And more than a third of those citations were occurring *after* the paper had been officially retracted and after the retraction notice had been published. Health care providers are being taught to rely on this form of evidence when making decisions, but the information community is struggling to account for retracted materials within that methodology. Bakker provided other examples that are included in the full paper based on this session.

12.4.

Hannah Heckner, Product Strategist at Silverchair, raised three questions:

(1) What are the intervention points for stopping the spread of retractions?
(2) Which gatekeepers can intervene and/or disseminate retraction status?
(3) What are the impediments to open access dissemination of retraction statuses and retraction notices?

She really did not answer the questions, but spoke generally and noted that there is a lot of inconsistency when you look at publisher sites as to how they communicate the retracted status of an article. There is the possibility to provide watermarking on the actual front end, be it on the article PDF in addition to the actual article page. This is something that is adopted by many publishers, but certainly there is room for improvement. Past that front end, she believes that there is a lot of opportunity in increasing the metadata vocabulary around retracted research and opportunities to add new tags to articles to communicate about retractions, perhaps even a sub-vocabulary where the publisher/platform provider can talk about the types of retractions. She suggested that the various article artifacts be made more open on platforms. She believes that a move towards more posting of open data, posting of article versions, even posting of more information about the life cycle of an article, would be helpful to increase the transparency about research and just perhaps this would shine light on retractions before they became a larger problem.

12.5.

Jodi Schneider, Assistant Professor, University of Illinois Urbana-Champaign, gave a summary of the recommendations from the project for **R**educing the **I**nadvertent **S**pread of **R**etracted **S**cience [67] (RISRS2020). They are drafting a white paper and currently have five top-level recommendations. First, to make retraction information easy to find and use. Second, to recommend retraction metadata and a taxonomy of retraction statuses that can be adopted by stakeholders. Third, to develop best practices for coordinating the retraction process. Fourth, to educate and socialize researchers and the public about retraction and post-publication stewardship of the scientific record. And fifth, to develop standard software and databases to support sustainable data quality. She said that the group would welcome input.

12.6. Preprint platforms and fake news

The final speaker in this session was Michele Avissar-Whiting, Editor-in-Chief of the Research Square [68] Preprint platform (where the RISRS report has been posted and is in revision). She opened her brief comments with a quote from the author Marchette Chute, who said, “Nothing can so quickly blur and distort the facts as desire, the wish to use the facts for some purpose of your own. And nothing can so surely destroy the truth [69]”. She said that is has been fascinating to see how people use information and to think about the challenge that these behaviors pose to researchers, journalists, and the platforms such as Research Square that have been hosting early outputs of information in the form of preprints. She said that most preprint servers or platforms are not “anything goes” platforms. They do filter out submissions that are clearly pseudoscientific, ethically dubious, or potentially dangerous. But they do not routinely block the posting of papers based on methodological flaws, poor or opaque reporting, or specious conclusions. Preprint servers are already not totally passive hosts for research. And the last few months have taught us that we may be able to play a more active role in ensuring that people, at minimum, don’t come away with totally misguided ideas about what a study means.

She went on to give some examples of how people have used the information in preprints to reinforce their own agenda - especially in the news. They are included in the full paper that is based on this session. She said that the whole pandemic has been a trial by fire for those whose job it is to think about the role of preprints, how they’re being received, and establishing policies around them, etc. As a preprint server, she feels that it is incumbent on them not only to screen out bad material and include disclaimers for everything else, but also to take other actions within their means to provide clues to the rigor of the study and to help people make sense of it. And these are features that have the potential to add value above and beyond what a standard editorial or peer review process can offer.

She then showed a Pentateuch framework [70] that is used to measurer scientific rigor, which was introduced by Casadevall and Fang. It lists the five components by which to determine to what extent a given study can be trusted: (1) intellectual honesty; (2) probability and statistics; (3) logic; (4) experimental redundancy; and (5) error analysis. And Research Square uses the framework for evaluation at the preprint stage, before or alongside a standard peer review process. They now use public-facing badges so that users of their preprint server are alerted to potential problems. You can see them if you go to the Research Square website.

As I have noted throughout this summary, an excellent paper based upon this session appears elsewhere in this issue of *Information Services and Use.*

13. The CRediT taxonomy

There was a session on the values and challenges of the CRediT Taxonomy [71] - a taxonomy with which I was unfamiliar. It is a high-level taxonomy that includes fourteen roles that can be used to represent the roles typically played by contributors to collaborative research that has scientific scholarly output. The roles describe each contributor's specific contribution to the scholarly outputs.

I asked one of the panelists, Alex Holcombe, a professor of psychology at the University of Sydney, to submit a manuscript which he did, and that introduced me to another new term "tenzing [72]", about which I knew nothing. The premise of the paper is that information about the people associated with a published journal article has been traditionally handled manually and unsystematically. However, as large-scale collaboration, sometimes referred to as "team science", is now common, a more structured and easy-to-automate approach to managing meta-data is required. In his paper he describes how the latest version of tenzing combined with the CTediT Taxonomy helps researchers collect and structure contributor information efficiently and without frustration. I am not a technical person so I will not even attempt to explain it - just read the article that appears elsewhere in this issue of *Information Services and Use*.

14. KBART phase III: Unresolved questions

During the "NISO update" session at the conference, members of the KBART (Knowledge Base and Related Tools) [73] Standing Committee presented their plans for the development of KBART Phase III, a revision of the KBART Recommended Practice. As an FYI, KBART is a NISO Recommended Practice that facilitates the transfer of holdings metadata from content providers to knowledge base suppliers and libraries. Knowledge bases are widely-used to support library link resolvers and electronic resource management systems (ERMs) [74].

A key problem is that librarians often do not know which package they need to link to their system and therefore often do not always "turn on" all the files for which they have paid. At last year's conference Lola Estelle, Digital Library Specialist, at SPIE, said that packages appear differently across discovery systems and that platform migration can cause problems. Selecting and enabling the correct package within an ERM, discovery layer, or link resolver is difficult because there are multiple similarly-named packages. Librarians seek guidance from vendors, but vendors themselves often do not know which package to enable and may not have a way to see the selection screens in ERMs and related systems. She said that content providers who wish to provide guidance around knowledge base content selection must ensure that their KBART files are properly named and that they provide sufficient documentation and training for librarians.

KBART is about to undergo a revision and because of its significant importance to Electronic Management Systems, the KBART Subcommittee sought input from conference attendees on how KBART is currently being used and what new content types it should support in the future. Approximately one hundred and forty people attended the main session and they provided extensive input which is detailed in a paper entitled "KBART Phase III: Unresolved Questions" that appears elsewhere in this issue of *Information Services and Use*. If you are interested in KBART it is a must read to see what the subcommittee is grappling with and how the attendees would like to see the Practice upgraded. You can give your input as well because contact information is provided.

15. Closing

In his closing comments, Todd Carpenter, NISO Executive Director, noted that he had no realistic expectations that the conference would turn out as well as it did. He believes that the success of this program was a direct result from setting out the global team in advance who would be willing to - and did - engage an international audience. He added that in 2020, two hundred and forty people attended the conference in Baltimore, which was, from his perspective, a great turnout. But he had no reason to believe that the 2021 conference would bring together eight hundred and thirty-five people across twenty-six countries. He said that he hopes to build on the ideas that emerged from this conference - to take those ideas and turn them into projects so that NISO can continue to engage and work towards a world where everyone has access, unfettered access to information.

16. Conclusion

As you can see from this overview, there was no major theme to the conference other then it being a global conversation. Having said that, there were common themes throughout and some of them resonated even with the topics of the prior year's conference.

- Creating rich metadata is essential to facilitate information discovery and preservation.
- Citing and reusing datasets requires a cultural and behavioral shift among researchers.
- Ensuring that datasets used for Machine Learning and Predictive Analytics are complete, unbiased, and relevant to the project at hand is absolutely essential to quality output and diligent attention to this challenge is required.
- Ensuring that XML is consistent and aligned with industry standards and using persistent identifiers for software and data is key to the proper linking of data citations and software citations so that they get counted and the creators are given credit.
- Using standards is essential to the global sharing of data and scholarly information.

The majority of the presentations that I "attended" were excellent. I was especially impressed with the presentation on misinformation and fake news. Clearly, so much can result in the spread of incorrect data, especially if that data supports a personal bias of the reader and I was impressed with what Research Square has implemented to ensure that the preprints on their server have badges to ensure that readers are made aware of the potential shortfalls of manuscripts. I always like it when (1) I learn about new things such as "tenzing" or the AI projects that Andromeda Yelton described, or China's new Personal Information Protection Law that will be implemented in November; (2) when a presentation can make me look at something such as the GDPR and view it in a new light; or (3) when I am made aware of an issue to which I had never given thought such as the challenges facing the preservation of Indigenous Knowledge. Those are the take-aways that, for me, make attending a conference worthwhile and those are the things that made attending the 2021 NISO Plus conference worthwhile for me.

Last year Todd Carpenter called the 2020 conference a “Grand Experiment”. When writing the conclusion of my conference overview I honestly said that I believed that the experiment was successful. I also said, that as a chemist, I am quite familiar with experiments and am used to tweaking them to improve results. And as successful as the 2020 meeting was, in my opinion it needed tweaking, and to some extent the 2021conference reflected positive modifications. But I still believe that there needs to be more of the information industry thought-leadership that NFAIS conferences offered. Perhaps the next global conversation can include some specific information industry issues and show how they are being handled around the world. I would like to know how various countries are dealing with the “ownership” of the outputs from AI projects such as the images and paintings mentioned by Kaufman. I would like to know how publishers around the globe are providing guidelines to authors regarding data and software citations. A “global conversation” on specific industry-wide issues would be both fascinating and informative.

Having said that, I congratulate the NISO team and their conference planning committee on pulling together an excellent virtual conference. From my perspective, it has been the best virtual conference that I have attended throughout the Pandemic - technically flawless and well-executed. NISO should publish a Best Practice on virtual conferences and make it a global standard!

My congratulations to Todd and his team for a job well done!!

Additional Information

The NISO 2022 Conference [75] will take place completely virtually from February 15-17, 2022, and registration is now open.

If permission was given to post them, the speaker slides that were used during the 2021 NISO Plus Conference are freely-accessible in the repository on the NISO Plus 2021website. The same applies to the session recordings. To access them go to the website [76], scroll down and click on “draft Program”. The complete program is there and if you click on “NISO Repository” you can access the slides; for the videos, click on “NISO Videos”. I do not know how long they will be available.

About the Author: Bonnie Lawlor served from 2002 through December 2013 as the Executive Director of the National Federation of Advanced Information Services (NFAIS), an international membership organization comprised of the world’s leading content and information technology providers. She is currently an NFAIS Honorary Fellow. She is a Fellow and active member of the American Chemical Society and an active member the International Union of Pure and Applied Chemistry for which she chairs the Subcommittee on Publications. She is also on the Board of the Philosopher’s Information Center, the producer of the *Philosopher’s Index,* and she serves as a member of the Editorial Advisory Board for *Information Services and Use.*

About NISO: NISO, the National Information Standards Organization, is a non-profit association accredited by the American National Standards Institute (ANSI). It identifies, develops, maintains, and publishes technical standards and recommended practices to manage information in today’s continually changing digital environment. NISO standards apply to both traditional and new technologies and to information across its whole lifecycle, from creation through documentation, use, repurposing, storage, metadata, and preservation.

Founded in 1939, incorporated as a not-for-profit education association in 1983, and assuming its current name the following year, NISO draws its support from the communities that is serves. The leaders of about one hundred organizations in the fields of publishing, libraries, IT, and media serve as its Voting

Members. More than five hundred experts and practitioners from across the information community serve on NISO working groups, committees, and as officers of the association.

Throughout the year NISO offers a cutting-edge educational program focused on current standards issues and workshops on emerging topics, which often lead to the formation of committees to develop new standards. NISO recognizes that standards must reflect global needs and that our community is increasingly interconnected and international. Designated by ANSI to represent U.S. interests as the Technical Advisory Group (TAG) to the International Organization for Standardization's (ISO) Technical Committee 46 on Information and Documentation. NISO also serves as the Secretariat for Subcommittee 9 on Identification and Description, with its Executive Director, Todd Carpenter, serving as the SC 9 Secretary.

References

[1] NISO Plus 2021Content Now Openly Available, Press Release, August 26, 2021, https://niso.plus/niso-plus-2021-content-now-openly-available/, accessed September 29, 2021.

[2] Cory Doctorow, Wikipedia, https://en.wikipedia.org/wiki/cory_doctorow, accessed September 26, 2021.

[3] Surveillance Capitalism, Wikipedia, https://en.wikipedia.org/wiki/surveillance_capitalism, accessed September 27, 2021.

[4] The Social Dilemma, Netflix, https://www.imdb.com/title/tt11464826/, accessed September 27, 2021.

[5] J. Wihbey, Facebook experiment in social influence and political mobilization, *The Journalist's Resource* (2012), https://journalistsresource.org/politics-and-government/facebook-61-million-person-experiment-social-influence-political-mobilization/, accessed September 27, 2021.

[6] Network Effects, Wikipedia, https://en.wikipedia.org/wiki/Network_effect, accessed September 29, 2021.

[7] https://www.thebookseller.com/news/cma-clears-prh-ss-sale-1259480, accessed September 27, 2021.

[8] https://en.wikipedia.org/wiki/Robert_Bork, accessed September 27, 2021.

[9] B. Kepes, Google users - you're the product, not the customer, *Forbes* (2013), https://www.forbes.com/sites/benkepes/2013/12/04/google-users-youre-the-product-not-the-customer/?sh=20f5210976d6, accessed September 29, 2021.

[10] https://www.dimensions.ai, accessed September 30, 2021.

[11] http://digital-science.com, accessed September 30, 2021.

[12] https://www.dimensions.ai/webinars/dimensions-on-google-bigquery/, accessed September 29, 2021.

[13] The diagram can be accessed at: https://www.nature.com/immersive/d42859-019-00121-0/index.html, accessed September 30, 2021.

[14] https://datacite.org, accessed September 30, 2021.

[15] https://orcid.org, accessed September 30, 2021.

[16] https://ror.org, accessed September 30, 2021.

[17] https://grid.ac, accessed September 30, 2021.

[18] https://project-freya.eu.en, accessed September 30, 2021.

[19] https://www.gbif.org, accessed September 30, 2021.

[20] https://crossref.org, accessed September 30, 2021.

[21] *Standardized Markup for Journal Articles: Journal Article Tag Suite*, NISO, https://www.niso.org/standards-committees/jats, accessed October 3, 2021.

[22] https://www.force11.org/group/software-citation-working-group, accessed October 3, 2021.

[23] B. Lawlor, An overview of the 2020 NISO Plus inaugural annual conference: *A Grand Experiment*, *Information Services and Use* **40**(3) (2020), 145, https://content.iospress.com/journals/information-services-and-use/40/3, accessed October 3, 2021.

[24] https://csir-forig.org.gh, accessed October 3, 2021.

[25] S.S. Rao, Indigenous knowledge organization: an Indian scenario, *International Journal of Information Management* **26**(3) (2006), 224–233, https://www.sciencedirect.com/science/article/pii/S0268401206000089, accessed August 11, 2021.

[26] https://www.copyright.com, accessed October 3, 2021.

[27] https://www.eff.org/deeplinks/2014/06/another-fair-use-victory-book-scanning-hathitrust, accessed October 3, 2021; and https://www.wired.com/2013/11/google-2/, accessed October 3, 2021.

[28] https://h2o.law.harvard.edu/cases/5141, accessed October 6, 2021.

[29] https://law.justia.com/cases/federal/appellate-courts/ca2/15-3885/15-3885-2018-02-27.html, accessed October 6, 2021.

[30] D. Boyd, Questioning the legitimacy of data, *Information Services and Use* **40**(3) (2020), 259–272, IOS Press, https://content.iospress.com/articles/information-services-and-use/isu200098, accessed October 6, 2021.

[31] https://www.theartstory.org/artist/prince-richard/, accessed October 21, 2021.

[32] https://en.wikipedia.org/wiki/Jeff_Koons, accessed October 21, 2021.

[33] For examples of machine-generated art: Machine Creativity Beats Some Modern Art, *MIT Technology Review*, June 20, 2017, https://www.technologyreview.com/2017/06/30/150666/machine-creativity-beats-some-modern-art/, accessed October 2, 2021.

[34] Who owns Artificial Intelligence Created Art?", https://www.sybariscollection.com/owns-artificial-intelligence-created-art-copyright/, accessed October 6, 2021.

[35] https://ischool.sjsu.edu, accessed October 28, 2021.

[36] https://github.com/cmoa/teenie-week-of-play, accessed October 109, 2021.

[37] https://cmoa.org/art/Teenie-Harris-archive/, accessed October 10, 2021.

[38] https://transkribus.eu/lite/, accessed October 10, 2021; and https://readcoop.eu/transkribus/, accessed October 10, 2021.

[39] https://laesekompas.dk, accessed October 10, 2021 (the site does not see to have an English option).

[40] https://hamlet.andromedayelton.com, accessed October 10, 2021.

[41] https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html, accessed October 10, 2021.

[42] https://dhlab.yale.edu/projects/pixplot/, accessed October 10, 2021.

[43] https://citizen-dj.labs.loc.gov, accessed October 10, 2021.

[44] https://www.media.mit.edu/projects/gender-shades/overview/, accessed October 10, 2021.

[45] *Describing Archives: A Content Standard DACS 2019.0.3*, Society of American Archivists, 2020, https://mysaa.archivists.org/productdetails?id=a1B5a00000heUDGEA2, accessed October 11, 2021. This version is currently undergoing revision.

[46] L. Parsons, Harvard alumna amanda gorman delivered a souring inauguration poem, *Harvard Gazette* (2021), https://news.Harvard.edu/gazette/story/2021/01/amanda-gormans-inauguration-poem-the-hill-we-climb/, accessed October 11, 2021.

[47] https://arctoris.com, accessed October 21, 2021.

[48] *AI Ethics in Scholarly Communication: STM Best Practice Principles for Ethical, Trustworthy and Human-centric AI*, STM, April 2021, https://www.stm-assoc.org/2021_05_11_STM_AI_White_Paper_April2021.pdf, accessed October 12, 2021.

[49] Tired of books written by authors? Try Booksby.ai - *Browse the bookshop for printed paperback books entirely generated by Artificial Intelligence*, https://booksby.ai, accessed October 13, 2021.

[50] M. Urberg, Digital humanities and standards: Let's get this conversation started, *Information Services and Use* **40**(3): https://content.iospress.com/journals/information-services-and-use/40/3, accessed October 16, 2021.

[51] https://en.wikipedia.org/wiki/Jisc, accessed October 13, 2021.

[52] The General Data Protection Regulation, 2016, https://en.wikipedia.org/wiki/General_Data_Protection_Regulation, accessed October 13, 2021.

[53] https://www.oag.ca.gov/privacy/ccpa, accessed October 14, 2021.

[54] https://ag.ny.gov/internet/data-breach, accessed October 14, 2021.

[55] https://www.cdc.gov/phlp/publications/topic/hipaa.html, accessed October 14, 2021.

[56] https://theconceptofconsent.wordpress.com/continuous-consent/, accessed October 14, 2021.

[57] https://www.fieldfisher.com/en/insights/china's-personal-information-protection-law-what-d. accessed October 16, 2021.

[58] https://assets.kpmg/content/dam/kpmg/cn/pdf/en/2017/02/overview-of-cybersecurity-law.pdf, accessed October 16, 2021.

[59] https://www.orrick.com/en/insights/2021/09/Chinas-New-Data-Security-Law-What-International-Companies-Need-to-Know, accessed October 16, 2021.

[60] https://www.niso.org/node/25942, accessed October 14, 2021.

[61] NISO RP-19-2020 Open Discovery Initiative: Promoting Transparency in Discovery, https://www.niso.org/standards-committees/odi, accessed October 15, 2021.

[62] https://www.projectcounter.org, accessed October 16, 2021.

[63] C.B. Anderson, Preservation and archiving of digital media, *Information Services and Use* **30**(3) 201–208: https://content.iospress.com/journals/information-services-and-use/40/3, accessed October 16, 2021.

[64] https://www.opscidia.com, accessed October 17, 2021.

[65] Open Access, Wikipedia, https://en.wikipedia.org/wiki/Open_Access, accessed October 17, 2021.

[66] https://www.vietsch-foundation.org/projects/, accessed October 17, 2021.

[67] Schneider, J., “Recommendations from the RISRS Report: Reducing the Inadvertent Spread of Retracted Science,” preprint in revision, Research Square, https://www.researchsquare.com/article.rs-783543.v1, accessed October 17, 2021.
[68] https://www.researchsquare.com. Accessed October 17, 2021.
[69] https://todayinsci.com/C/Chute_Marchette/ChuteMarchette-Quotations.htm, accessed October 17, 2021.
[70] A. Casadevall and F.C. Fang, Rigorous science: A how-to guide, *Europe PMC* **7**(6) (2016), https://Europepmc.org/article/MED/27834205, accessed October 17, 2021.
[71] https://credit.niso.org, accessed October 17, 2021.
[72] A.O. Holcombe, M. Kovacs, F. Aust and B. Aczel, Documenting contributions to scholarly articles using CRediT and tenzing, *PLOS One* **15**(12) (2020); https://pubmed.ncbi.nlm.nih.gov/33383578, accessed October 17, 2021.
[73] https://www.niso.org/standards-committees/kbart, accessed October 17, 2021.
[74] https://www.niso.org/standards-committees/kbart, accessed October 17, 2021.
[75] https://www.niso.org/niso-io/2021/08/all-about-niso-plus-2022, accessed October 20, 2021.
[76] https://niso.org/events/2021/02/niso-plus-2021, accessed October 20, 2021.